Algorithmic Representation of Visual Information

Daby Sow and Alexandros Eleftheriadis
Department of Electrical Engineering
Columbia University
New York, NY, 10027, USA
{daby,eleft}@ee.columbia.edu

Abstract

In [6], we have introduced Complexity Distortion Theory, a mathematical framework characterizing the design of programmable communication systems. In this paper, we show how Complexity Distortion Theory fits in the MPEG-4 context and more generally in any system allowing programmability, by formalizing the concept of programmable decoders. We also show how it can be used to design intelligent encoders at two flexibility levels: the first one corresponding to the case where flexibility in the algorithm selection is allowed and the second where downloadability of new tools for representation is also allowed.

1 Introduction

From the developments of Information and Rate Distortion Theories introduced in the late 40's by C. E. Shannon, emerged several coding techniques commonly used in many standards for representation of information. These techniques evolved out of the desire to design optimal representations in a compression sense, with little flexibility, for specific information sources. Modern applications are now hosted on sophisticated general-purpose computers; furthermore, compression is just one of many desirable characteristics. Features such as flexibility in algorithm selection and even downloadability of new tools, are quickly becoming fundamental requirements for new image and video representation approaches. This problem was first addressed by the MPEG-2 standard with the definition of profiles and scalability modes. The next logical extention is the substitution of the traditional decoding device by a computer. This general trend resulted in the use of programming languages in the ongoing MPEG-4 standardization effort. Such a modification makes the resulting communication system too general to be analyzed by the classical Information and Rate Distortion Theories. To make use of all the potential provided by these and other algorithmic approaches, a significant amount of intelligence must be added at both the decoding and encoding ends of the communication system. In [6], we have introduced Complexity Distortion Theory, a mathematical framework by which such a communication system can be characterized.

In this paper, our goal is to show how Complexity Distortion Theory fits in the MPEG-4 context and more generally in any system allowing programmability, by formalizing the concept of programmable decoders and how it can be used to design intelligent encoders at two flexibility levels: the first one corresponding to the case where flexibility in the algorithm selection is allowed and the second where downloadability of new tools for representation is also allowed.

2 Complexity Distortion Theory

Complexity Distortion Theory [6] uses an algorithmic approach to analyze and unify all coding techniques, from classical entropy methods to more elaborate schemes: fractal-based, model-based, etc. As shown in Figure 1, the key aspect of this theory is the substitution of Shannon's classical communication system by Chaitin's model where the decoder is a universal Turing machine [2]. Codewords are now programs. In such a framework, the algorithm becomes itself the content rather than just a method used to process the content. In contrast with traditional complexity theory, we also allow lossy representations, a more appropriate way to study the performance of practical coding algorithms for audio visual objects. Information is measured by the Kolmogorov-Chaitin complexity where programs are required to be prefix free, an important property for communication problems where instantaneous codes governed by the Kraft inequality are always desired. In fact it is proved in [2] that it is possible to construct what is called a Chaitin computer, a machine with programs satisfying the prefix free condition.

Definition 1 *[2] The Kolmogorov-Chaitin complexity $K_{\mathcal{U}}(x_n)$ of a string x_n is the minimum length over all prefix free programs written for a universal Turing machine $\mathcal{U}$ that print x_n and halt.*
If the knowledge of a string y is provided to $\mathcal{U}$, we assume that the shortest description of y is also provided to $\mathcal{U}$. The conditional Kolmogorov-Chaitin complexity $K_{\mathcal{U}}(x_n \mid y)$ is then the minimum length over all prefix free programs written for $\mathcal{U}$ that print x_n and halt, when the knowledge of y is provided.

Since all computers are equivalent up to a constant c measuring the length of the simulation that must be run by a computer to simulate the actions of another computer, the subscript identifying the computer used will be dropped in the rest of the paper.

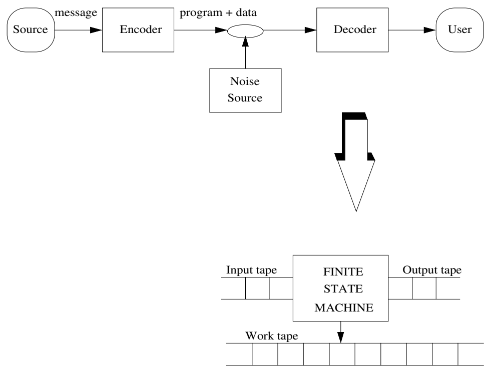

Figure 1: Chaitin's Communication System.

Theorem 1 *[2] The symmetry of information:*

$$K(x) - K(x \mid y) = K(y) - K(y \mid x) + O(1)$$

In [3] it is shown that for ergodic sources there are some equivalences between Shannon's entropy rate H_{rate} of the process and $K(x_n)$ the complexity of an observation of the process: $\lim_{n\to\infty} \frac{K(x_n)}{n} = \lim_{n\to\infty} \frac{K(x_n)}{n} = H_{rate}$. We have extended this result to the lossy case. A similar work is proposed in [8]. Here, we adopt a more intuitive approach, based on the notion of types, which we expect will be able to give us a better framework to tackle both theoretical and practical problems.

Definition 2 *[6] Let D be a distortion measure according to any valid distance metric $d(\cdot,\cdot)$. On a given object x, with D as a constraint, we introduce distortion in order to minimize the complexity of the resulting object y. If we have more than one object y with the same optimal complexity, we select the closest to x. If many objects x are equidistant to x, we arbitrarily select one of them and map it to x. We clearly define a function from the set of source objects x to the set of distorted objects y. We denote this function $\mathcal{D}$.*
We also define $D_{max} = \sup_{D:\ d(x_n,\mathcal{D}(x_n))>0} D$.

Definition 3 *[6] The complexity distortion function is:*

$$C(D) = \lim_{n\to\infty} E[\frac{K(\mathcal{D}(x_n))}{n}]$$

Using symmetry of information and noting that $\lim_{n\to\infty} \frac{K(\mathcal{D}(x_n)|x_n)}{n} = 0$, we get:

$$C(D) = \lim_{n\to\infty} E[\frac{K(x_n) - K(x_n \mid \mathcal{D}x_n)}{n}]$$

since $\lim_{n\to\infty} \frac{K(x_n)}{n}$ and $\lim_{n\to\infty} \frac{K(x_n|\mathcal{D}(x_n))}{n}$ are bounded by 1.

Theorem 2 *[6] For ergodic sources,*

$$C(D) = \lim_{n\to\infty} \frac{K(x_n) - K(x_n \mid \mathcal{D}(x_n))}{n} = R(D)$$

with probability 1, $R(D)$ being the rate distortion function of the source defined as: $R(D) = \lim_{n\to\infty} \inf_{Q(x_n \mid y_n)} \frac{I(x_n,y_n)}{n}$

Outline of the Proof: Let $y = \mathcal{D}(x_n)$. The proof can be divided in two main parts:
In part 1, we define two programs yielding upper bounds for the $\lim_{n\to\infty} \frac{K(x_n)}{n}$ and $\lim_{n\to\infty} \frac{K(x_n|y)}{n}$. To define and measure the length of these programs, we use the concept of type or empirical distribution.

Definition 4 *The type $P_x^n(\underline{i},\underline{j})$ of a sequence of n symbols $\underline{i}$ observed in an observation x_k of k symbols is the relative proportion of occurrence of $\underline{i}$ in x_k.*

For an ergodic source, if $k \to \infty$ the type converges to the probability distribution of $\underline{i}$. We call P^∞ the ergodic type. If $n \to \infty$ but with $O(n) < O(k)$, we get the type of an infinite observation x_n of the sample path x_k

Definition 5 *Let $\mathcal{P}$ be the set of all ergodic types for observations of length n and constant type P. The type class of P denoted $T_x(P)$ is:*

$$T_x(P) = \{x\epsilon\mathcal{H}^n : P_x^n = P\},$$

$\mathcal{H}$ being the source alphabet. The conditional type class of P denoted $T_y(P)$ is:

$$T_y(P) = \{x\epsilon S_y : P_x^n = P\},$$

S_y being the set of all sequences x_n with optimal distortion y.

Using the Shannon-McMillan-Breiman theorem, we get the following lemma:

Lemma 1 *When n goes to infinity,*

$$\mid T_x(P) \mid \leq 2^{nH_{rate}}, \ \mid T_y(P) \mid = \lim_{n\to\infty} \frac{1}{Q(x_n \mid y)}$$

Q being the probability distribution of the source.

We are now ready to define the programs:
When no prior knowledge is available (description of x_n), do:
- From all the possible sequences in $T_x(P)$, output the i^{th} corresponding to the one we want to describe.
When the knowledge of $y = \mathcal{D}(x_n)$ is available (description of $x_n \mid y$), do:
- From all the possible sequences in $T_y(P)$, output the i'^{th} corresponding to the one we want to describe.
These program have a length equal to a constant plus the size in bits of the indices i and i', which are respectively equal to the size in bit of the cardinal of the type class and the size in bit of the cardinal of

the conditional type class. Therefore, we conclude the first part of the proof with:

$$\lim_{n\to\infty} \frac{K(x_n)}{n} \leq H_{rate}$$

$$\lim_{n\to\infty} \frac{K(x_n \mid y)}{n} \leq \lim_{n\to\infty} \frac{-\log Q(x_n \mid y)}{n} = \frac{\log \mid T_y \mid}{n}$$

In part 2, we use the incompressible property of infinite random strings, proposed in [2] to show that the set of strings from the source with a complexity smaller than the length of the previous programs has measure zero according to the probability distribution of the source.

Theorem 3 *[2] For each fixed constant r, the number of object x_n with $K(x_n) \leq n+K(n)-r$ does not exceed $2^{n-r+O(1)}$*

Applying this theorem we show that:

$$\lim_{n\to\infty} Q(K(x_n) < nH_{rate}) \leq \lim_{n\to\infty} \frac{1}{2^{K(n)+O(1)}} = 0$$

Similarly,

$$\lim_{n\to\infty} Q(K(x_n) \mid y) < \log(\mid T_y \mid) \leq \lim_{n\to\infty} \mathcal{C}.Q(x_n)$$

$$\mathcal{C} = \mid \{x_n \epsilon T_x(P) : K(x_n \mid y) < \log(\mid T_y \mid\} \mid$$

Obviously,

$$\lim_{n\to\infty} \mathcal{C} \leq \lim_{n\to\infty} 2^{\log|T_y(P)|} = \lim_{n\to\infty} \frac{1}{Q(x_n \mid y)}$$

Therefore,

$$\lim_{n\to\infty} Q(K(x_n) \mid y) < \log(\mid T_y \mid) \leq \lim_{n\to\infty} Q(y) = 0,$$

if $0 < D < D_{max}$. Using the definition of $\mathcal{D}$ and the subadditive ergodic theorem, it remains to argue that $\lim_{n\to\infty} \frac{K(x_n|\mathcal{D}(x_n))}{n}$ is a constant. Also, since $\mathcal{D}$ is defined in order to minimize the complexity of x_n under a fidelity criterion, we reach the rate distortion function.

$\square$

Because of the Halting problem,the Kolmogorov-Chaitin complexity is not Turing computable. Like Shannon's probabilistic measure of information, this measure requires an infinite amount of computational resources. But in this case, resource bounds can be introduced naturally, something that is difficult to do in the traditional information theoretic context. The first resource constraint is space. The decoding device has only finitely many finite input and working tapes. In other words the memory space is finite. Consequently, the encoder must segment the information into pieces that can not only fit in the decoder memory but also do not require too much space on the working tapes during the computation. Such a segmentation has been studied in [9] but in the traditional information theoretic context. Extension to the second type of resource bounds, time, is difficult to do without the substitution of the decoding device by a Turing machine. In fact, putting some time limit on the computation of the decoder removes the disturbing Halting problem and yields a recursive resource bound complexity which is an excellent practical indicator of the performance of the communication system. Such considerations are necessary for the design of practical systems.

3 Modern Visual Representation Systems

Using the Chruch-Turing thesis[1], it is easy to see that Complexity Distortion Theory provides the utmost flexibility and universality in terms of capabilities of the receiver by including the estimation of statistics in the coding problem. Application of traditional information theory in practice is limited by the probabilistic assumptions on the source. In the classical framework, research on universal coding provides partial answers to this problem by performing a search for a good stochastic model in the set of stationary ergodic sources. Models that do not require any probabilistic assumptions on the source such as fractal coding, mode-based-coding, are excluded from the search space. This results in a conceptual separation between today's practical codec systems and the classical theory. Via the Church-Turing thesis, Complexity Distortion Theory provides a reconcilliation between practice and theory for any coding scheme. In general, the task of the encoder consists on finding a good model in which the observation of the source fits well. To this model, some parameters specific to the observation must be added to describe the observation. For example, to represent a Sierpinski triangle, a simple iterated system model with 3 affine transformations can be used. The description on how this system has to be used constitutes the model. The values of the coefficients of the transformations constitute the data part of the representation. This separation between model and data is underlined in Figure 1. There is an important tradeoff between the two. The model extracts the regularities in the observation. On one hand, a specific observation can be compactly described with the help of a complex model for it. On the other hand, with a very simple model, the representation is less specific to the observed data and it can be used to represent a broader class of observations. In this case, less regularities are squeezed out of the data. Following the Occam Razor principle, the best model is the simplest one and when compression is the main issue, it corresponds to the association model data with total length equal to the Kolmogorov-Chaitin complexity. Due to the non computability of this measure of information the quest for the global optimal point in a compression sense is hopeless. Even with resource bounds, we

[1] The Church-Turing thesis states that the class of algorithmically computable numerical functions (in the intuitive sense) coincides with the class of partial recursive functions (computed by Turing machines)

may run into computation difficulties because of the exponential character of the large set of candidates. It is wiser to try to approximate the solution. The quest for the optimal solution is studied at two levels of flexibility.

At the first level, we reduce the complexity of the problem and the flexibility of the codec by limiting the model search space to a finite number of algorithms. The encoder faces a detection problem. It is its responsibility to index the right model and transmit this information along with the data characterizing a particular audio-visual object. By enlarging the term complexity to any mode of description associated with a particular encoding scheme, different complexities defined like in [5] can be associated with classes of models defined by different coding schemes. Consequently, the search for a good model is organized by partitioning the model search space into smaller classes where we are looking for the simplest description depending on the application. When compression is the main issue, this detection problem is linked to the Minimum Description Length approaches (MDL) which estimate statistics based on the length of short descriptions [5]. In [7], links between MDL and Bayesianism reasoning are investigated using the Kolmogorov-Chaitin complexity. Some equivalences between the two are shown using the concept of universal distribution derived from Solomonoff's induction theory which is also based on short program lengths. These observations show how Complexity Distortion Theory can be used to analyze this problem and provide answers to the design of good optimal encoders at this flexibility level.

At the second level, the encoder is free to use the full capabilities of the decoding Turing machine. Downloadability of new tools is allowed. Due to the non computability of the Kolmogorov-Chaitin complexity, finding the optimal combination model data in a compression sense is impossible. This observation does not contradict classical information theoretic results when we assume that the statistics of the source are unknown. The difficulty is primarily due to the exponential size of the set of possible models. Even with resource bounds, an exhaustive search would run into obvious computational difficulties. This problem has been studied in [4]. Approximation of the solution for representation of images are given using Genetic Algorithms. At the first step, an initial finite set of programs are tested. Depending on their performances, fitness values are attributed to these programs. Using a natural selection principle, the best programs are kept, combined and mutated probabilistically in order to generate new "child" programs still syntactically correct. After several generations, this process gives an approximation of the solution. Such a programmatic approaches constitutes a new interesting way to represent information.

4 Concluding Remarks

In this paper, we have addressed the question of universal representation of audio-visual information using an algorithmic approach with a Turing machine at the decoding end of the communication system. We gave a brief outline of the equivalence between Complexity Distortion theory, a new perspective for media representation and the classical Rate Distortion Theory. This equivalence allows to generalize the classical concept of information based on probabilities and provide a unification of all coding techniques, classical and modern, under a single theory. We then addressed the design of modern visual representation systems at two levels of flexibility. At the first level, the representation tools are standardized and the encoder faces a detection task. At the second level, full flexibility is available and, because of the non computability of the Kolmogorov-Chaitin complexity shortest descriptions can only be approximated.

References

[1] O. Avaro, P. Chou, A. Eleftheriadis, C. Herpel, and C. Reader, "The MPEG-4 System Description Language," *Signal Processing: Image Communication, Special Issue on MPEG-4*, Vol. 9, No 4, pp. 385-432, July 1997.

[2] G. J. Chaitin, "A Theory of Program Size Formally Identical to Information Theory," *Journal of the Association for Computing Machinery*, Vol. 22, No 3. pp. 329-340, July 1975.

[3] A. K. Leung-Yan-Cheong and T. Cover, "Some Equivalences between Shannon's Entropy and Kolmogorov Complexity," *IEEE Transactions on Information Theory*, Vol. 20, pp. 331-339, 1978.

[4] P. Nordin and W. Banzhaf, "Programmatic Compression of Images and Sound," *Proceedings of the First Annual Conference on Genetic Programming.* J. R. Koza, D. E. Goldberg, D. B. Fogel and R. L. Riolo ed., The MIT Press, pp. 345-350, July 28-31 1996.

[5] J. Rissanen, "Stochastic Complexity in Statistical Inquiry," *World Scientific*, 1989.

[6] D. Sow and A. Eleftheriadis, "Complexity Distortion Theory," *Proceedings 1997 IEEE International Symposium on Information Theory, Ulm Germany*, pp. 188, June 29 - July 4 1997.

[7] P. Vitanyi and M. Li, "Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity," *Submitted to: IEEE Transactions on Information Theory*, September 1996.

[8] E. H. Yang and S. Y. Shen, "Distortion Program-Size Complexity with Respect to Fidelity Criterion and Rate-Distortion Function," *Transactions on Information Theory*, Vol. 39, No 1, pp. 288-292, January 1993.

[9] J. Ziv, "Back from Infinity: A Constrained Resources Approach to Information Theory," *Proceedings 1997 IEEE International Symposium on Information Theory, Ulm Germany*, pp. 4, June 29 - July 4 1997.