Language and Information Content

By R. C. PUETTER[1]

[1]Center for Astrophysics and Space Sciences,
University of California, San Diego
9500 Gilman Drive
La Jolla, CA, 92093-0111, USA
rpuetterucsd.edu

This lecture discusses the role of language and information theory concepts for data compression and solving the inverse problem. The concept of Algorithmic Information Content (AIC) is introduced and shown to be crucial to achieving optimal data compression and optimized Bayesian priors for image reconstruction. The dependence of the AIC on the selection of language then suggests how efficient coordinate systems for the inverse problem may be selected. This motivates the selection of a multiresolution language for the reconstruction of generic images.

1. Introduction

The many problems that arise with classical image reconstruction, e.g. Goodness-of-Fit (GOF) methods (= Maximum-Likelihood methods), and Maximum Entropy (ME) image reconstruction, all stem from the reconstructed image being too complex. This is seen in the severe over-fitting of the data by these methods and in their production of spurious sources. Indeed, the basic improvement of ME methods over GOF methods results directly from attempting to reduce the amount of information contained in the reconstructed image—one attempts to maximize the entropy, or disorder in the image. This strongly suggests that controlling the image information content is a key issue for successful image reconstruction. That this is true should be especially apparent when the collected data has associated measurement noise. In this case the data contains much more detailed information than is relevant to the image—e.g. the precise values of the pixels in a digital image are unimportant since noise contributes to their value. Hence, it is clear that there should be some smaller, but quantifiable, amount of information in the image, and that the ability to identify this information and ignore the other details of the data would be valuable. Other clues that controlling the information content of the image is highly desireable come directly from Bayesian estimation theory. Here, the Bayesian prior always favors the simplest (and hence the *a priori* most likely) hypothesis (=image or image/model) for explaining the data. This is the Bayesian embodiment of Occam's Razor.

It is the goal of this lecture to find a suitable method for controlling the information content of the image and to use this to constrain the image reconstruction process. Looking to the information sciences, we find that there is a rich literature on the nature of information and the role of language in the specification of information. Futhermore, from the field of image processing, we find that multiresolution languages and operators provide powerful tools for describing the structure of generic images. These insights will allow us to develop concise and powerful languages to handle generic images and develop highly optimized Bayesian priors for our image reconstruction applications. The concepts crucial to these goals are described below.

2. Types of Information

The concepts of information and entropy have been discussed in numerous works. While the thermodynamic concepts of entropy have a very specific meaning and apply to the state of an ensemble of particles, the concepts of information have a much broader scope. Nonetheless, the laws of statistical physics can be derived directly from information theory concepts (Jaynes 1957; Jaynes 1963; Hobson 1971), and information and entropy (or negentropy) are often talked about as the same quantities. The most common definition of information, now known as Shannon information, was introduced in 1948 (Shannon 1948; Shannon and Weaver 1949) and specifies the average information contained in a string of symbols transmitted across a communication line. In this definition, the information per symbol is given by

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad , \tag{2.1}$$

where n is the total number of possible symbols, and p_i is the probability of occurance of the i^{th} symbol. There are many reasons why the Shannon information is a suitable definition for information. One of the most convincing arguments is given by the *Noiseless Source Coding Theorem* (see, for example, Rabbani and Jones 1991), which states that for a ergodic source with an alphabet of n characters, that as the size of the transmitted string approaches infinite length, it is possible to construct a unique random length binary code for the string that has an average length in bits per encoded symbol equal to, but no less than, the Shannon entropy.

While the *Noiseless Source Coding Theorem* in and of itself speaks strongly for Shannon entropy as a sensible information measure, there are still other motivations. For example it can be proven that the Shannon information is the unique and self-consistent measure which satifies conditions which would reasonably be expected of a measure of information (Khinchin 1957; Feinstein 1958; Hobson 1971). Using Khinchin's formulation of the properties uniquely defining an information measure, it is found that only the function H satisfies the conditions that (1) H takes it's largest value when all the p_i are equal, (2) H doesn't change it's value when additional impossible events are added—i.e. events for which $p_i = 0$, and (3) the information in two not necessarily independent events is given by the information of the first event plus the expectation value of the additional information provided by the second type of event after the first event has occurred, i.e. if A and B are sets of events, then

$$H(AB) = H(A) + \sum_i p_i H(B|A_i) \quad , \tag{2.2}$$

$$H(AB) = H(B) + \sum_j p_j H(A|B_j) \quad , \tag{2.3}$$

where $H(X|Y)$ is the additional information in X given that Y has occurred.

Readers who have had classical training in physics will be interested that the formula of equation (2.1) is essentially equivalent to the physical definition of entropy, $\sigma = k \ln W$, (modulo a constant to convert logarithm types) where k is Boltzman's constant and W is the total number of available states. In fact, equation (2.1) is the Boltzman (or Gibbs) definition of entropy. Kittel (Kittel 1969—see page 118) gives a particularly fluid derivation of the Boltzman definition of entropy starting from the Boltzman factor, i.e. $p_i = Z^{-1} \exp(-E_i/kT)$. Thus Shannon information is seen to be not merely a mathematical construct of theoretical interest, but has a substantial practical "bite"—statistical physics describes the real world and works exceedingly well.

While we have now provided motivation for the merits of Shannon information, of what use is this concept for solving the inverse problem in general or image reconstruction in particular. To be sure, ME methods make use of the entropy or Shannon information definition. However, Shannon information doesn't get to the heart of the problem. This definition of information doesn't allow us to describe the *information content* of an image or data set. As has been pointed out by several authors (Wicken 1987; Chaiten 1982), Shannon information is an ensemble notion. It is a measure of ignorance concerning possible realizations of events which have a given *a priori* probability distribution, and measures the statistical rarity or "surprise value" of a particular realization (Cherry 1978). It does not deal with the information content expressed by a given realization. For this we need a different concept, i.e. that of algorithmic information content (AIC), algorithmic randomness, or algorithmic complexity (Solomonoff 1964; Kolmogorov 1965; Chaiten 1966). As commonly used in information or computer science, the AIC of a string of characters is defined to be the size of the minimum computer program required for a universal computer to produce the specified string as its output. (A computer U is universal if for any other computer M there is a prefix program μ such that the program μp makes U perform the same tasks as performed by computer M running the program p, i.e. the program μ makes computer U simulate computer M.) As suggested by the name, AIC describes the information content of the specific item under scrutiny. In fact it does this is a very practical way, i.e. algorithmically. It provides a prescription for how the data can be reproduced. It should also be clear that the AIC represents the optimal compression of the data. There is no shorter description that will allow a complete reconstruction of the original information. In fact, it is this optimal compression of the information that will allow us to optimally extract an image in the image reconstruction problem in a practical manner.

3. The Role of Language in AIC and Bayesian Estimation

The term "randomness" comes into the definition of AIC since random strings of characters have maximum complexity or information content. This is because if the pattern of the string is non-random, there is some way of describing the string which is typically shorter than listing the string itself. For example, in strings of 1s and 0s, a string with 10^9 copies of the number 1 would take a lot of paper to write down, but can be described with a single sentence. Similarly this same string with a 0 placed in the millionth place can also easily be described without wasting paper. However, a totally random string of 10^9 digits cannot be so easily described, and if no rule is known (and this is what we mean by random), only the actual string will impart all of the information. It is well known that quantification of the AIC is a function of the "richness" of the language used to describe the character string or data set. In the example used above, we demonstrated how the English language could be used to briefly describe large strings of digits. Other languages, e.g. the binary digits 0 and 1, decimal digits, or even hexidecimal digits, do not provide such a terse description. An example of the character strings used to express a given number in various languages is given in Table 1.

The above discussion makes it quite obvious that rich, i.e complex, languages allow a terse discription of a data set—the AIC is low for rich languages. This seems like an obvious advantage. It says that the data can be compressed to a higher degree and that the information can be more precisely located and identified. Similar concepts occur in Bayesian estimation, e.g. the concept of Occam's Razor. Clearly, in the Bayesian estimation problem, it is highly beneficial to develop concise, simple hypotheses (images

Character Expressions of a Number

Character String	Language
11111010011100111011	binary
3723473	octal
102581	decimal
FA73B	hexadecimal

TABLE 1. The character representation of a given number in different languages. The simplest language, i.e. binary, with two characters in the language, requires 20 characters to express the number. The richest language, hexadecimal, with 16 characters in the language, requires only 5 characters to express the number.

or image/model pairs) to explain the data. Such simple hypotheses give rise to optimized priors [$p(I|M)$ or $p(I|M)p(M)$—see Lecture 1 of this series]. Consequently, they produce optimal reconstructions and a more likely M.A.P. image or image/model. Thus it is seen that minimizing the AIC of the image/model, optimally compressing the information in the data, and performing highly optimized Bayesian image reconstruction are directly related. Hence the goal for improved image reconstruction is clear. One must improve the language in which the reconstruction is performed, or equivalently the language in which the image is described. The language must be rich so that the description can be concise, the AIC low, and the Bayesian prior optimized. The fact that the description will be concise suggests that the language for describing the image must be "natural" in some sense, and we turn now to a discussion of possible languages for image description.

4. Languages for Images

There are many "languages" for images in current use. The most common of these are usually not even thought of as languages since they seem so natural. For example, in electronic or digital imaging we essentially expect to see every image in the form of a rectangular pixel array. This is so natural that we don't even consider the consequences of this language for the image. However there are consequences. The pixels are almost always rectangular and so is the picture format. However this may not be optimal for image reconstruction. Certainly in the case of astronomical imaging, many objects are not square, e.g. stars. Also images of star clusters or galaxies, etc., seldom uniformly fill a rectagular picture to its corners. Hence these clearly are not an optimal format or pixel shape for astronomical situations. There are usually too many pixels in the image and many pixels are used to describe single objects such as stars.

Another common basis for images is the Fourier basis. This, of course, is quite natural in the case of radio interferometry where the collected data are actually points in the Fourier domain. This basis is often also used when image deconvolution is attempted for high signal-to-noise (SNR) data—see Lecture 1 of this series. Astronomers also often think in terms of the Fourier basis when considering the fundamental limits on spatial frequency response of a telescope, e.g. the sharp cut off in frequency space giving rise to the telescope's Airy diffraction pattern. However, there are many reasons why the Fourier basis often is inappropriate (and even misleading) for astronomical imaging. This is seen quite easily once one realizes that most objects in astronomical images are spatially localized and different objects usually bear no causal relationship with each other. By contrast, each Fourier component of an image contains information from all parts of the image. Hence every location in the image becomes intertwined in the Fourier transform. This is quite unnatural and is at the root of the difficulties associated with Fourier deconvolution in the moderate to low SNR case—noise from all over the image gets mixed together.

Thinking about images in the Fourier domain can also be misleading. A common misunderstanding, for example, is that the finite size of a telescope prevents the detection of star pairs more closely separated than the full-width-half-max (FWHM) of the diffraction pattern. This is simply not true. If the SNR of the data is high enough one can easily tell the difference between a single Airy pattern and the sum of two closely spaced Airy patterns. What is really happening is that the Airy pattern prevents information on the Fourier components with frequencies higher than a certain cut off frequency. This does not mean, however, that high spatial resolution imaging is prevented. What it means is that there is some ambiguity in the image. One can add, for example, any image that has only components with frequencies higher than the cut-off frequency and not change the collected data. Usually, however, images such as this are unphysical and can be discriminated against by other means, e.g. by probabilistic methods.

In addition to the above languages there are two more languages that should be discussed and which are in many ways more suitable that either the pixel or Fourier bases. They are the wavelet basis and the pixon basis. Both of these languages are largely motivated by the structure of picture information. Indeed, the while the pixel and Fourier bases are extreme in their view point (one being very local, the other being completely global), the wavelet and pixon bases take a middle-ground approach. Before discussing these bases further, let us first ask the question: "What is the nature of picture information?" This, of course, can have many answers, and the answer changes, depending on the object at which the photographer points his camera. We are interested, here, in "generic" images. So let use first ask the general question: "What do we really expect when we take a picture?" Thinking about this in the broadest sense, all we can really say is that we expect the picture to contain some limited amount of information. We expect, in particular, that the manufacturer of the film or the designer of the camera will have made the grain or pixel size fine enough to do service to the picture, i.e. not to leave out any important detail. We expect interesting features, perhaps people, with open spaces between them. In other words, we expect that at each point in the image there is a finest spatial scale of interest and that there is no information content below this scale. Indeed, this is how photographic grain sizes are chosen and why data is not sampled finer than the Nyquist frequency when pixel elements are at a premium.

How does one capture this prior expectation in mathematic form and incorporate it into a set of image basis functions? Again, the key comes from thinking about photographic grain sizes or Nyquist sampling. We would do just as well at recording the picture information with large photographic grains in portions of the image with coarse structure. We need only have fine grains when we need to record fine spatial structure. This means that the picture information can be dealt with by using variable sized cells, with the cell sizes set so as to capture the spatial information present. This is the fundamental idea behind wavelets and multiresolution techniques in general. These approaches seek to localize the information in different parts of the image. Wavelets attempt to localize the frequency information, while multiresolution techniques seek to analyze or localize the information on different spatial scales.

Having had the foregoing dissussion, we are ready to discuss in more detail both wavelets, multiresolution methods, and our current implementation of pixon-based methods. Wavelets are basically an extension of Fourier methods in which the frequency information is localized. A complete description of wavelet theory is beyond the scope of this lecture, but suffice it to say that wavelets are a linear, orthogonal set of basis functions that do a much better job (are more concise) of describing images that contain localized structure. However as we have shown for the reconstruction problem, arguments based on optimization of the Bayesian prior indicate that the languages that are the most concise will provide the best reconstruction. So we are left with the question is the wavelet basis the most concise language for describing generic images. The Fourier basis failed because images with highly localized structure and large empty spaces have Fourier components at all frequencies and so are very verbose. Pixel bases fail because they may use many more pixels than necessary to describe a large smooth structure in the image. Wavelet bases fall somewhere in the middle. They localize the information. Yet they still describe the localized information in terms of a number of oscillating basis functions. High frequency oscillating functions are required to describe small scale structure and many components are required to "cancel-out" the resulting wavelet oscillations at the edges of fine bumps.

Like the wavelet approach, multiresolution techniques attempt to quantify the information present in an image at a variety of scales. Unlike wavelets, however, these techniques do not require the basis functions to be linear or orthogonal. They simply try to quantify the information. Our pixon-based methods currently use a multiresolution language. We would like to stress here that our concept of a pixon—see Lecture 3—is that each pixon can be identified with a unit of information (in the AIC sense) in the image, and that the collection of an image's pixons are the most concise description possible in the chosen language, i.e. the pixon basis is the AIC of the image. We have chosen to implement pixon-based methods using a multiresolution language since it seems that this language is "natural" for generic images—generic images have a variety of structure on a variety of spatial scales. It is clear from this that our pixon bases need not be either linear nor othogonal, since having these extra constraints on the selection of the pixon basis might compromise the all important goal of the basis being concise.

While we have now made it clear that a basis that allows concise descriptions is important, there is, of course, another important property of an image basis. It must be complete (or near complete) to be useful. This property is easily seen to be an attribute of the pixel, Fourier, and wavelet bases, at least for images that are definable on the pixel grid. We must be careful, however, to select a pixon basis that is complete. Since we have used a multiresolution language, completeness can be easily insured if single (or effectively single) pixels are included in the set of basis functions. This is clearly desirable anyway since one may wish to describe structures that are this fine in spatial scale.

5. Conclusions

This lecture has described the basic information properties in an image. We have pointed out that each image has a finite information content and that the information in we are interested is the Algorithmic Information Content as defined by the computer and information sciences. We argued that if we were able to indentify this information then we would be in a position to develop a scheme for its optimal extraction. We discussed the role of language in determining the AIC of a data set and the direct relationship of a concise (rich) language and small AIC values, optimized Bayesian priors, and consequently high quality solutions to the inverse problem. We then discussed appropriate languages for image description, stressing that the languages should be concise and complete. We finally argued that for generic images a multiresolution description should be superior to the direct pixel basis and the Fourier and wavelet bases. In the next lecture we shall see how to explicitly use the ideas developed in this lecture to construct practical pixon bases.

REFERENCES

CHAITIN, G. J. 1966, *J. Ass. Comput. Mach.* **13**, p. 547.

CHAITEN, G. J. 1982, Algorithmic Information Science, *Encyclopedia of Statistical Science* **1**, pp. 38-41. Wiley.

CHERRY, C. 1978, *On Human Communication,* 3rd edition. MIT Press.

FEINSTEIN, A. 1958, *Foundations of Information Theory.* McGraw-Hill.

HOBSON, A. 1971, *Concepts in Statistical Mechanics.* Gordon and Breach.

JAYNES, E. T. 1957, Information Theory and Statistical Mechanics, *Phys. Rev.***106**, pp. 171-190.

JANES, E. T. 1963, Information Theory and Statistical Mechanics, In *Statistical Physics*, ed. K. W. Ford, pp. 181-218. W. A. Benjamin, Inc.

KHINCHIN, A. I. 1957, *Mathematical Foundations of Information Theory.* Dover.

KITTEL, C. 1969, *Thermal Physics.* Wiley.

KOLMOGOROV, A. N. 1965, *Inf. Transmission* **1**, p. 3.

RABBANI, M., AND JONES, P. W. 1991, *Digital Image Compression Techniques.* SPIE Optical Engineering Press.

SHANNON, C. E. 1948, A Mathematical Theory of Communication, *Bell Systems Tecnical Journal* **27**, 379.

SHANNON, C. E., AND WEAVER, W. 1949, *The Mathematical Theory of Communication.* University of Illinois Press.

SOLOMONOFF, R. 1964, *Inf. Control* **7**, p. 1.

WICKEN, J. 1987, Entropy and Information: Suggestions for a Common Language, *Philos. Sci.* **54**, pp. 176-193.