

LAB 3 - Entropy Coding*Lab Report Due: February 16, 2006*

In image coding, we are essentially concerned with the problem of representation of an image (or sequence of images) with low bit rates (bits/pixel) for storage or bits/second when considering transmission. There are two basic approaches: predictive coding, carried out in the spatial domain and transform coding, that we may refer to as coding in the data domain or frequency domain respectively. (We will examine these two methods later.) Both these techniques are statistical in nature. (We do not consider here the concept of abstracting information in a “meaningful” way, such as abstracting features that the human observer may consider important.)

The basic statistical property upon which image compression techniques depend upon is inter-element correlation: To the extent to which element values in a locality in an image are similar, the magnitude of one may be estimated from the value(s) of one or more of the others nearby. High correlation implies a power spectral distribution that is strongly low pass in nature (agree?), thus requiring low coding capacity for transmission. Hence, the degree to which images may be compressed crucially depends on their correlation properties.

Since the purpose of image coding for storage and transmission is the retrieval or communication of information, it is logical to investigate the information content of the “source”. That is, to ask the question: What are the minimum number of bits per image pixel necessary to code an image so that it may be reconstructed without error? We will henceforth refer to this information preserving limit as *MNB*. Note that it *is indeed possible* to code at below these rates in return for an acceptable degree of distortion in reconstruction. But the fact remains that the *MNB* is still a good guide to limits when allowing specific degrees of distortion during reconstruction.

How do we determine the *MNB*? Well, that depends on the image model that we assume. There can be a large number of interrelationships between the occurrence of various element values within an image which may then present a complex image model. However, with certain simplifications, such as assuming low order distributions, it are not difficult to obtain an estimate. We begin with the first-order distribution.

1 First-order distribution

The fundamental notion in information theory is that of uncertainty. Unlikely events carry more information than likely ones. Consider an image where x represents element intensity values. Let the event $X = x$, where X is an arbitrary image element (a random variable)

and x is a particular luminance value occurring with probability $p(x)$. Then the “self-information” of X , $I(X)$ that occurrence is defined as

$$I(X) = \log[1/p(X)] = -\log p(X)$$

1.1 Self Information as *uncertainty or surprise*

The base of the logarithm is arbitrary, though it will of course, affect the numerical value of $I(X)$. Base 2 gives results in bits and base e in natural units or nats. As a simple example, we conclude that based on this definition, a binary word B bits long contains B bits of self information.

Note that this definition implies that (self) information contained is *large* when it is *unlikely* and small when there is certainty. $-\log_2(0.9) = 0.152$ and $-\log_2(0.1) = 3.322$ This makes sense: that when something is highly probable, there is no surprise (there is certainty) and hence (as per Shannon’s definition) little information. Conversely, when something with low probability occurs (such as winning the lottery) then there is surprise and hence (as per Shannon’s definition), more information. And if something is absolutely certain, there is no information and the information is zero (right)?

1.2 Shannon’s Entropy as Average Surprise, Average (self) Information

Consider an image (source) where pixel values assume 2^B values. Then the average self-information per picture element (over the whole image), is defined as

$$H(X) = \sum_{X=1}^{2^B} p(X)I(X) = - \sum_{X=1}^{2^B} p(X)\log_2 p(X) \quad \text{bits per pixel}$$

$H(X)$ is known as *entropy* or sometimes as *Shannon entropy* after the mathematician Claude Shannon. This entropy tells us, how surprised we will be, *on average* when we learn the value of the variable X .

Example:

Consider a random variable X that has the following distribution:

$$\begin{array}{ll} P(X=A) = & 1/2 \\ P(X=C) = & 1/8 \end{array} \quad \begin{array}{ll} P(X=B) = & 1/4 \\ P(X=D) = & 1/8 \end{array}$$

For this random variable X , its entropy is

$$H(X) = \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) + \frac{1}{4}(3) = 1.75$$

That is, the *average uncertainty* about this random variable is 1.75. What does that mean? How do we obtain a quantitative sense of 1.75? Well, consider the case of the random variable - roll of a loaded dice - that always turns up 6, i.e. $P(X = 6) = 1$ and of course $P(X = 1) = 0, P(X = 2) = 0$ etc. Here we easily see that $H(X)=0$. (We assume that $\log_2 0 = 0$). That is, average uncertainty is now 0 here. When do we very high uncertainty? Clearly, when all events are equally likely (nothing is certain). For the same random variable above, with all four events equally likely, we get $H(X) = 2$. So now average uncertainty is 2, compared with 1.75 when there was more certainty. So perhaps this provides a bit more sense of the concept of *average uncertainty*.

At a more physical level, average uncertainty can be related to the number of yes-no questions that it will take for you to figure out the value of X . That is, if your first guess is $X = A$, you would be right half the time. Thus, $1/2$ the time you would need only 1 question to guess correctly. If you guessed incorrectly, your next guess could be $X = B$. Again, you would be correct half of the time. As a result, $1/4$ of the time it would take you 2 guesses to determine X . Similarly, $1/4$ of the time you would need 3 guesses to determine X . Adding, we would get, the average number of guesses as

$$\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) = 1.75$$

which is of course what we got earlier as the value of $H(X)$. Thus, the idea of entropy as the average number of yes-no guesses is consonant with our earlier interpretation of entropy as a measure of uncertainty.

Note that it is *not* a function of X but a function of the pdf of X . Note that we could just as easily have used the index 0 to 2^{B-1} for the values of $H(X)$ depend solely on the probabilities of occurrences and *not* on their numerical values. We also note that $H(X) \geq 0$ (i.e. always psd) and that $H(X) = 0$ if and only if X is known with certainty (what type of pdf does that then imply?) Note also that there are several other ways to view Shannon's entropy as for example, entropy from a thermodynamic point of view.

1.3 Shannon’s Entropy as Coding

It is not difficult to show that the maximum value of $H(x)$ occurs when all outcomes are equally likely, in which case

$$H(x) = \sum_{x=1}^{2^B} 2^{-B} \log_2(2^{-B}) = B \text{ bits}$$

Hence, we say that the average number of bits per image element needed to represent the image array without loss of information is B . All naturally produced images have a pdf that departs from this uniform situation and hence can be represented by fewer than B bits on average. In particular then, you can always conclude that for a *wint8* image, we can of course code it using 8 bits. *However* knowing this 8 bits/pixel is the worst case (as per our previous argument) and knowing that the pdf will typically tend not to be uniform, we in theory, know that a lesser number of bits/pixel can be used for coding it. As a matter of fact, there exists a fairly simple method (Huffman) code to implement that limit. This is called “entropy based coding” and is not uncommonly applied. Note that, as per our earlier discussion, this is not optimal since (i) the pdf may not be uniform (ii) the concept does not make allowances for inter-element correlation. However, it is often applied *after* other coding techniques have been applied.

For a physical interpretation, one sees that entropy of a source refers to the *uncertainty* in the source. We saw that a uniform distribution maximizes entropy - as a matter of fact it corresponds to complete uncertainty since everything is equally likely to occur - you can’t get more uncertain than that!

2 Second-order distribution

The relationship developed in the previous section determines average information content of an image on a ‘per element’ basis, i.e. we have not given any consideration to other relations such as, interdependence between image data points, that may exist. We should expect to do considerably better by taking such interdependence into account. Images have many areas where the luminance values do not vary to any great extent. A logical continuation is to the previous development, then, is to consider elements *not singly, but in pairs* and to determine a corresponding value for the entropy of such an arrangement, which should, by the previous argument, be smaller than the per element entropy $H(X)$.

Accordingly, a joint entropy $H(X, Y)$ is defined in a similar way to $H(X)$, where the relevant event is $(X = x, Y = y)$ and the two equalities are to be simultaneously satisfied by X and

Y , which are the values of adjacent elements in the image array. (Note that the X and Y may be pairs taken horizontally or vertically). Thus,

$$H(X, Y) = - \sum_{X, Y=1}^{2^B} p(X, Y) \log_2 p(X, Y) \quad \text{bits} \quad (1)$$

is the average value of self-information of a pair of image elements. If there was *no* statistical relationship between the elements of any pair, then we would expect that $H(x, y) = 2H(x)$. In practice we will have

$$H(X, Y) \leq 2H(X) \quad (2)$$

and so

$$2H(X) - H(X, Y) \geq 0$$

and that reflects the fact that knowing the value of one element of a pair tells us, on average, something about its neighbor. Note that the qualification of ‘on average’ is important, since a pair of elements may be on either side of a sharp luminance transition of an image.

The difference between $H(X, Y)$ and $H(X)$ is

$$H(Y/X) = - \sum_{X=1}^{2^B} p(X, Y) \log_2 p(Y/X) = H(X, Y) - H(X) \quad (3)$$

where $H(Y/X)$ represents the average self-information in the occurrence of Y (i.e. $Y = y$ given that we know that the previous element has the value X , and is termed the conditional entropy of Y given X).

Two limiting cases may serve to clarify relationship between (2) and (3). Suppose first the (2) is satisfied with the equality. Then, from (3) we have

$$H(Y/X) = H(X) \quad (4)$$

Again, suppose that $H(y/x) = 0$. Then,

$$H(X, Y) = H(X). \quad (5)$$

Now, *both* elements values in the pair are completely specified by the single amount of information $H(x)$. Knowing x informs us of the exact value of y , or in other words, every pair of elements has members of the same value.

EXPERIMENT

1. Copy images.mat from <http://www.cems.uvm.edu/~mirchand/classes/EE276/Images/>. Work with any 3-4 images.
2. For, say, a 256x256 images, determine the number of bits/pixel presently being used to code the image.
3. Plot the distribution of a number of different types of images.
(You will find, for instance, that for low-detail images, you will often see two or three peaks.)
4. For a number of images, determine the first-order entropy and hence the minimum number of bits with which it could be coded (were such a coder available). You can use MATLAB's *entropy* and *entropyfilt* functions
5. Very briefly, what calculations would be required for the next higher order entropy?

LAB REPORT

1. Brief report of items 2,3,4 and 5.

Ref: Transform coding of images, R.J.Clarke. Academic Press, 1985

Class notes:mirchand-ee276 – February – 2006