Hierarchical Probability Estimation

Martin Bichsel and Krystyna W. Ohnesorge*
University of Zurich
Department of Computer Science
MultiMedia Laboratory
Winterthurerstr. 190
8057 Zurich
Switzerland
mbichsel@ifi.unizh.ch

Abstract

Estimating probabilities based on measured numbers of occurrences of events provides a central link from probability theory to real world applications. In an important class of applications the probabilistic events correspond to the digitized outcome of an analog sensor. This paper shows theoretically and experimentally that such events are governed by a natural similarity relation which imposes subtle but highly effective *a priori* constraints on the meta-probability distribution, *i.e.* the probability distribution of probability values. The application of these constraints significantly improves the probability estimates based on the measurements. The results are applied to estimating order-1 Markov model parameters in image processing applications.

1 Introduction

Virtually every real-world signal processing problem has a non-deterministic component. Therefore, its adequate description has to be statistical, in terms of joint probability distributions. Numerous examples can be found in many different areas such as pattern recognition, data compression, and process control.

Powerful techniques and theorems are at hand, allowing to find optimal solutions, as soon as the probability distributions are known. Examples are Bayes' rule [1] for pattern recognition problems and Shannon's theorem [2] for data compression. Difficulties, however, arise when it comes to determining the probability distribution of the admissible patterns. We can *never* measure these probabilities directly – we can only measure the number of occurrences of patterns in a statistical experiment and *estimate* the probabilities based on the measurements. The difficulties of statistical probability estimation can be illustrated with an example in which a coin is flipped twice and the outcome is twice head. Based on this experiment, there is no unique correct answer telling us the probability for head

*This work was supported by the consortium VISAGE, KWF grant No. 2440.1 (first author), and Swiss National Science Foundation grant #21-29915.90 (second author).

BMVC 1994 doi:10.5244/C.8.15

or tail. Of course, with any reasonable estimation method, the probability estimates usually get better the more experiments we make. Unfortunately, the coin example is quite typical for many situations in image analysis where the number of possible patterns, *i.e.* the number of different events, is comparable or even large compared to the number of measurements that can be taken with a reasonable effort. In such an experiment, most patterns occur only once or not at all.

Hence, the real, to a large extent unsolved, problem consists in estimating higher order joint probabilities of patterns, where a pattern is a group of data values taken from data points at specific relative positions. This paper derives an important constraint for data representing digitized outputs of an analog sensor. This constraint improves the joint and conditional probability estimates and is illustrated for probability distributions of patterns in natural images. The same considerations also hold for other data types such as audio data.

2 Exploiting hierarchy of digital sensor outputs

There are subtle differences between unrelated events on one side and events which are governed by a physically based similarity relation on the other side. In this section it will be shown that the two cases lead to different *a priori meta-probability* distributions (probability distributions of probability values) which lead to different relations between measured numbers of occurrences and probability estimates.

Let us carry out a classical statistical experiment, consisting of N trials, and count the number of occurrences n_i for each event $i \in \{1,...,m\}$ where the experiment has m distinguishable events. Based on these counts we want to estimate the probabilities $\{p\{1\},...,p\{m\}\}$. Let $\hat{p}\{i|\mathbf{n}\}$ denote such a probability estimate based on $\mathbf{n} = \{n_1,...,n_m\}$. This estimate denotes the expectation value $E\{\tilde{n}_i/\tilde{N}\}$ of measuring $\tilde{n}_i$ occurrences of event i in an *independent* (repeated) experiment consisting of $\tilde{N}$ trials, whereas the experiment is repeated under the same conditions that were used to determine the numbers n_i.

In the most unspecific form of a stochastic experiment, all events are equivalent and and we have no *a priori* expectation that would favor a particular combination of probability values. Equivalence means that index i is only a label with no further meaning. Equivalently, we could use any permutation of these indices. In this case the best probability estimate, based on the counts n_i, is

$$\hat{p}_e\{i|\mathbf{n}\} = \frac{n_i + 1}{N + m} \tag{1}$$

which is subsequently called *equivalence-based probability estimate.* This result is derived by applying a *principle of maximum ignorance* [15, 4] to the probability values $\mathbf{p} = \{p\{1\},...,p\{m\}\}$ which are treated as random variables. This principle of maximum ignorance is based on the *a priori* assumption that any combination of probability values fulfilling $\sum_{i=1}^{m} p\{i\} = 1$ is equally likely, *i.e.* $P\{\mathbf{p}\} = 1/V_{sm}$. The *meta-probability* distribution $P\{\mathbf{p}\}$ is the probability distribution of the random field $\mathbf{p}$ and V_{sm} is the volume of an m-dimensional unit simplex. The maximum ignorance principle is consistent with the assumption of equivalent random events since the meta-probability distribution $P\{p\{1\},...,p\{m\}\}$ is symmetric under a permutation of the indices.

Thus, as a rather surprising consequence of the maximum ignorance principle, the best estimate of $p\{i\}$ is $(n_i + 1)/(N + m)$ rather than n_i/N. The latter is subsequently called *direct probability estimate*. For $n_i \gg m$ and $N \gg m$ we have

$$\frac{n_i + 1}{N + m} = \frac{n_i}{N} \frac{(1 + \frac{1}{n_i})}{(1 + m/N)} = \frac{n_i}{N} \left(1 + \frac{1}{n_i}\right) \left(1 - \frac{m}{N} + O\left(\frac{m^2}{N^2}\right)\right) \qquad (2)$$

Therefore, Eq. 1 differs from n_i/N significantly for small values of N and n_i, only.

Now, let $I \in [0,..., M-1]$ denote a digital sensor output, *e.g.* a digitized pixel value of a CCD sensor, where M denotes the number of different digital values (typically M=256). This imposes a similarity relation onto the events which implies that I_3 is more different from I_1 than from I_2 if $I_1 < I_2 < I_3$ are three possible digital sensor outputs. Therefore, we may *a priori* expect that if $I_1 < I_2 < I_3$ then also $p\{I_3\}$ is likely to be more different from $p\{I_1\}$ than from $p\{I_2\}$, where the values $p\{I\}$ are again treated as random variables.

This expectation is based on the fact that a digital sensor value I corresponds to an analog variable which is the result of numerous conjoint and more or less correlated physical effects. Each sensor element (image pixel) of the image sensor in a camera receives its light from a particular surface element in the scene. Typical effects contributing to a pixel value are

- Surface properties such as albedo, roughness, and specularity
- Distribution of light falling onto the surface element(s)
- Orientation of the surface element(s)
- Noise of the camera and digitizer electronics, dust in the air, defocus

The continuous nature of macroscopic physical variables lets us assume that the probability distribution of each contributing effect is smooth and, therefore, the resulting probability distribution for the joint effect is also smooth.

In fact, uncorrelated Gaussian noise always causes the resulting probability distribution to be smooth on a scale given by the noise variance [5]. Similarly, each contributing effect imposes a more or less weak smoothness constraint on an appropriate scale. If $I_1 \approx I_2$ then many strong *a priori* constraints relate the probability value $p\{I_1\}$ and $p\{I_2\}$ to each other. If I_1 and I_2 are very different then only a few weak constraints apply. Thus, if the index I describes a sensor output, the permutation symmetry of the *a priori* meta-probability distribution $P\{\mathbf{p}\}$ is not expected to be valid. With this *a priori* knowledge, the principle of maximum ignorance which led to Eq. 1 is invalid. Hence, better probability estimation methods must exist.

A possible strategy to derive better probability estimates would consist in replacing the maximum ignorance principle by a different explicit *a priori* meta-probability distribution $P\{\mathbf{p}\}$. So far, unfortunately, we could no find a computationally manageable way to do this. We found, however, an efficient method that implicitly uses an improved meta-probability distribution by calculating the probabilities in a hierarchical way. This method mimics the hierarchical way in which the various effects play together to produce the digital sensor output.

The idea is to hierarchically subdivide the set of the M possible digital sensor outputs in pairs of subsets so that, for each division, the *a priori* knowledge about the probability of being within either subset is minimum. To each pair of subsets

we then apply the principle of maximum ignorance. For simplicity reasons it is assumed that M is a power of 2.

Let us illustrate this procedure for the example M=256. The procedure starts by dividing the complete set $\{0,..., 255\}$ of all digital sensor outputs into two subsets. Maximum ignorance is achieved if the number of elements in the two subsets are equal and the values in the first subset are maximally different from the values in the second subset. The best choice splits the complete set into the two subsets $\{0,..., 127\}$ and $\{128,..., 255\}$ which differ in the most significant bit. This splitting maximizes the mean squared difference as well as the mean absolute difference of the two sets. Although a few elements in the first set are still similar to elements in the second set, most elements are different. As a consequence, we have little *a priori* knowledge about the probabilities that a sensor value falls into either set so that the maximum ignorance principle is a good approximation to this situation.

Let us therefore treat this situation as an experiment with two events for which Eq. 1 applies. The two events are $I \in \{0,...,127\}$ and $I \in \{128,...,255\}$. Let h_k denote the number of occurrences of the sensor value k, where $k \in \{0,...,255\}$. Thus, $n_1 = \sum_{k=0}^{127} h_k$, $n_2 = \sum_{k=128}^{255} h_k$, m=2, and $N = n_1 + n_2$. Applying the maximum ignorance principle leads to $p\{I \in \{0,..., 127\}|\mathbf{h}\} = (1 + \sum_{k=0}^{127} h_k)/(2 + \sum_{k=0}^{255} h_k)$ and $p\{I \in \{128,..., 255\}|\mathbf{h}\} = (1 + \sum_{k=128}^{255} h_k)/(2 + \sum_{k=0}^{255} h_k)$ where $\mathbf{h}$ denotes $\{h_0,..., h_{255}\}$.

Given that the sensor value lies within either subset the procedure continues by again splitting each such that maximum ignorance is achieved for the two probabilities that the sensor value falls into either sub-subset. Repeating the previous arguments leads to a splitting into the two groups of sensor elements which differ in the second-most significant bit. Without loss of generality let us illustrate this for the set $\{0,..., 127\}$ which, thus, is subdivided into the sets $\{0,..., 63\}$ and $\{64,..., 127\}$. This division again makes the values in the first set maximally different from the values in the second so that the preceding arguments can be repeated and the principle of maximum ignorance is applied again. This leads to

$$\begin{aligned} p\{I \in \{\ 0,...,\ 63\}|I \in \{0,..., 127\}, \mathbf{h}\} &= \left(1 + \sum_{k=0}^{63} h_k\right) / \left(2 + \sum_{k=0}^{127} h_k\right) \\ p\{I \in \{64,..., 127\}|I \in \{0,..., 127\}, \mathbf{h}\} &= \left(1 + \sum_{k=64}^{127} h_k\right) / \left(2 + \sum_{k=0}^{127} h_k\right) \end{aligned} \quad (3)$$

where $p\{I \in \{0,..., 63\}|I \in \{0,..., 127\}, \mathbf{h}\}$, *e.g.*, describes the conditional probability that the value falls into the lower quarter given that it is within the lower half and given the measurements $\mathbf{h}$. Combining this with the previous results leads to

$$\begin{aligned} p\{I \in \{\ 0,...,\ 63\}|\mathbf{h}\} &= p\{I \in \{\ 0,...,\ 63\}|I \in \{0,...,127\}, \mathbf{h}\} \cdot p\{I \in \{0,...,127\}|\mathbf{h}\} \\ p\{I \in \{64,...,127\}|\mathbf{h}\} &= p\{I \in \{64,...,127\}|I \in \{0,...,127\}, \mathbf{h}\} \cdot p\{I \in \{0,...,127\}|\mathbf{h}\} \end{aligned} \quad (4)$$

The *a priori* expectation that the probability $p\{I \in \{0,..., 63\}|\mathbf{h}\}$ is more similar to $p\{I \in \{64,..., 127\}|\mathbf{h}\}$ than to $p\{I \in \{128,..., 191\}|\mathbf{h}\}$ is implicitly taken account for since the former two probability estimates have a common factor. The first factor is calculated with higher precision and has a smaller tendency towards 1/2 than the second factor, due to the fact that the first factor contains more summands. This is an additional effect which implicitly takes account of the *a priori* expectation.

The above splitting procedure is repeated by further subdividing each subset until each subset consists only of one element $\{I\}$. In the case $M = 256$, the final probability $p\{I|\mathbf{h}\}$ then consists of a product with 8 terms.

In order to be able to express this idea in a compact form let as define the hierarchical sets $\mathcal{S}_k(I)$ of all digital sensor outputs which differ from I at most in the k least significant bits.

$$\mathcal{S}_k(I) = \{2^k \cdot (I \div 2^k), ..., 2^k \cdot ((I \div 2^k) + 1) - 1\} \tag{5}$$

where $\div$ denotes integer division. In the example k=2, i=5 we obtain $\mathcal{S}_2(5)$={4,...,7} which are all values that differ from 5 at most in the two least significant bits.

With this definition, the hierarchical probability estimate $\hat{p}_h\{I|\mathbf{h}\}$ for the value I is defined:

$$\hat{p}_h\{I|\mathbf{h}\} = \Pi_{k=0}^{log_2(M)-1} \frac{(\sum_{j \in \mathcal{S}_k(I)} h_j) + 1}{(\sum_{j \in \mathcal{S}_{k+1}(I)} h_j) + 2} \tag{6}$$

This hierarchical probability estimate, also, has the desired property that for large h_i and $H(\mathbf{h}) = \sum_0^{M-1} h_i$ it converges towards $h_I/H(\mathbf{h})$ since

$$\Pi_{k=0}^{log_2(M)-1} \frac{(\sum_{j \in \mathcal{S}_k(I)} h_j)}{(\sum_{j \in \mathcal{S}_{k+1}(I)} h_j)} = h_I/H(\mathbf{h}) \tag{7}$$

where $\mathcal{S}_0(I) = \{I\}$, $\mathcal{S}_{\log_2(M)}(I) = \{0,..., M-1\}$, and all intermediate terms cancel.

Equation 6 shows that hierarchical probability estimates can be calculated in a simple and efficient way. Compared to equivalence-based probability estimation (of the form $\hat{p}_e\{I|\mathbf{h}\} = (h_I + 1)/(H(\mathbf{h}) + M)$) it has the following advantages:

1. The maximum ignorance principle is applied to pairs of groups of sensory values for which the *a priori* smoothness expectation is low and hence the maximum ignorance principle is more appropriate. Therefore each term in Eq. 6 is closer to a best estimate.
2. For (most) data values that are close together, the terms with high k in Eq. 6 are identical and only the terms with low k differ. This reflects the *a priori* expectation that neighbouring values have similar probabilities.
3. The terms with low k, for which the *a priori* smoothness expectation is highest, are calculated with the least precision and show the largest tendency towards a probability 1/2. This is due to the fact that these terms have less summands than the terms with high k.

Although hierarchical probability estimates are not optimum, the following sections demonstrate that hierarchical probability estimation leads to a considerable improvement compared to traditional equivalence-based or direct estimation.

3 Testing the estimated probabilities

The quality of the estimated probabilities was tested with a number of experiments which investigate the predictability of joint statistics in natural images. The results are particularly important for adaptive image compression.

A general procedure for testing the estimated probabilities consists of carrying out two subsequent statistical experiments that are governed by the same statistics. In order to test the predictability of joint statistics in natural images according to

this general scheme, image are divided into two distinct parts. A joint statistics is collected in each part. Assuming a stationary joint statistics, the statistics in the second image part is then predicted based on the first part.

The measured numbers of occurrences in the first part are denoted by h_i, the measured numbers in the second part are denoted by $\tilde{h}_i$. Let $H(\mathbf{h}) = \sum_{i=0}^{M-1} h_i$ and $\tilde{H}(\tilde{\mathbf{h}}) = \sum_{i=0}^{M-1} \tilde{h}_i$ denote the total number of measurements in the two parts. Based on $\mathbf{h}$, the relative number of occurrences $\tilde{h}_i / \tilde{H}(\tilde{\mathbf{h}})$ should be predicted as accurately as possible.

The quality of the estimated probabilities is tested with two different methods. The first method calculates the squared deviation between the estimated probabilities in the first and $\tilde{h}_i / \tilde{H}(\tilde{\mathbf{h}})$ in the second part:

$$\tilde{\sigma}_e^2(\tilde{\mathbf{h}}, \hat{\mathbf{p}}) = \sum_{i=0}^{M-1} (\hat{p}\{i|\mathbf{h}\} - \tilde{h}_i / \tilde{H}(\tilde{\mathbf{h}}))^2 \tag{8}$$

This quantity is subsequently called *experimental variance.*

The second method makes use of Shannon's theorem which states that the minimum expected code length is achieved if and only if we assign exactly $-\log_2(p\{i\})$ bits to each symbol $i \in \{0..M-1\}$ where, in the context of image processing, i refers to a single gray value or a pattern consisting of two or more gray values. For the Shannon optimum, the mean code length per pattern is given by the entropy S:

$$S(\mathbf{p}) = -\sum_{i=0}^{M-1} p\{i\} \log_2(p\{i\}) \tag{9}$$

and Shannon's theorem states:

$$S(\mathbf{p}) \leq S_e(\mathbf{p} : \hat{\mathbf{p}}) := -\sum_{i=0}^{M-1} p\{i\} \log_2(\hat{p}\{i\}), \qquad \forall \hat{p}\{i\} \text{ with } \sum_{i=0}^{M-1} \hat{p}\{i\} = 1 \tag{10}$$

Let us identify $\hat{p}\{i\}$ with the estimated probabilities and use $p\{i\} = E(\tilde{h}_i / \tilde{H}(\tilde{\mathbf{h}}))$, where $\tilde{h}_i$ is the number of occurrences of pattern i in an *independent* test series with $\tilde{H}(\tilde{\mathbf{h}})$ examples. Thus, the quantity

$$\tilde{S}_e(\tilde{\mathbf{h}} : \hat{\mathbf{p}}) = -\sum_{i=0}^{M-1} \frac{\tilde{h}_i}{\tilde{H}(\tilde{\mathbf{h}})} \log_2(\hat{p}\{i\}) \tag{11}$$

which is subsequently called *experimental entropy,* is well suited to test the estimated probabilities $\hat{p}\{i\}$. The experimental entropy $\tilde{S}_e(\tilde{\mathbf{h}} : \hat{\mathbf{p}})$ describes the mean code length per pattern which an ideal lossless compression algorithm would achieve for the independent test data, based on the estimated probabilities $\hat{p}\{i\}$. The experimental entropy takes a minimum expectation value if the estimated probabilities and the true probabilities coincide. Experimental entropy is closely related to Kullback-Leibler information (directed divergence) [8]

$$\tilde{I}_e(\tilde{\mathbf{h}} : \hat{\mathbf{p}}) := \sum_{i=0}^{M-1} \frac{\tilde{h}_i}{\tilde{H}(\tilde{\mathbf{h}})} \log_2 \left(\frac{\tilde{h}_i / \tilde{H}(\tilde{\mathbf{h}})}{\hat{p}\{i\}} \right) \tag{12}$$

and we have $\tilde{S}_e(\tilde{\mathbf{h}} : \hat{\mathbf{p}}) = \tilde{I}_e(\tilde{\mathbf{h}} : \hat{\mathbf{p}}) - S(\tilde{\mathbf{h}})$ where $S(\tilde{\mathbf{h}}) = \sum \tilde{\mathbf{h}} / \tilde{H}(\tilde{\mathbf{h}}) \log_2(\tilde{\mathbf{h}} / \tilde{H}(\tilde{\mathbf{h}}))$ is a constant which is independent of the probability estimation method. Kullback-Leibler information is widely used in image processing for measuring the similarity of probability distributions.

4 Application to order-1 Markov models

We assume that the pattern formation process is ergodic and stationary [1], so that counting the number of occurrences of patterns in a sliding sub-window provides useful information about the global image statistics. Assuming, additionally, that the probability of a particular pixel only depends on its neighbourhood directly leads to Markov models.

Markov models are important signal processing tools. The parameters of a Markov model are conditional probabilities relating the value of a particular data point to the values of neighbouring points. Markov models are used, for example, in texture analysis [9], image compression [10], and speech recognition [11].

We verified the predictions of Section 2 for an order-1 model which is characterized by the conditional probabilities $p\{I(k+1)|I(k)\}$. In image processing, $I(k)$ denotes the value of pixel k, where the pixels are enumerated in scan-line order. For Markov models, $p\{I(k+1)|I(k)\}$ is independent of k so that we can write $p\{I_1|I_0\}$, instead.

Let M denote the number of gray levels. For each of the M contexts I_0, the model is characterized by M probability values so that the results of Sections 2 and 3 can be applied. By weighting the quality measures of Section 3 (Eqs. 8 and 11) with the relative number of occurrences of each context the following generalized quality measures are obtained:

$$\tilde{\sigma}_g^2(\tilde{\mathbf{h}}, \hat{\mathbf{p}}) = \sum_{I_0=0}^{M-1} \frac{\tilde{H}(I_0)}{\tilde{H}_{tot}} \sum_{I_1=0}^{M-1} \left(\hat{p}\{I_1|I_0, \mathbf{h}(I_0)\} - \tilde{h}(I_1, I_0)/\tilde{H}(I_0) \right)^2 \tag{13}$$

$$\begin{aligned} \tilde{S}_g(\tilde{\mathbf{h}}:\hat{\mathbf{p}}) &= -\sum_{I_0=0}^{M-1} \frac{\tilde{H}(I_0)}{\tilde{H}_{tot}} \sum_{I_1=0}^{M-1} \frac{\tilde{h}(I_1, I_0)}{\tilde{H}(I_0)} \log_2(\hat{p}\{I_1|I_0, \mathbf{h}(I_0)\}) \\ &= -\sum_{I_0=0}^{M-1} \sum_{I_1=0}^{M-1} \frac{\tilde{h}(I_1, I_0)}{\tilde{H}_{tot}} \log_2(\hat{p}\{I_1|I_0, \mathbf{h}(I_0)\}) \end{aligned} \tag{14}$$

In these formulas, $\tilde{h}(I_1, I_0)$ denotes the number of occurrences of the value I_1, given that the previous value is I_0. $\tilde{H}(I_0) = \sum_{I_1=0}^{M-1} \tilde{h}(I_1, I_0)$ is the number of occurrences of the value I_0, $\tilde{H}_{tot} = \sum_{I_0=0}^{M-1} \tilde{H}(I_0)$ is the total number of pixel pairs (I_1, I_0), and we write $\mathbf{h}(I_0) = \{h(0, I_0), ..., h(M-1, I_0)\}$.

The numbers $h(I_1, I_0)$ and $\tilde{h}(I_1, I_0)$ were calculated for a large number of images. In each image, $h(I_1, I_0)$ was calculated in the first quarter of the image, whereas $\tilde{h}(I_1, I_0)$ was calculated in the subsequent three quarters of the image. In this way, a large number of independent pairs of statistics were obtained.

For all such pairs, we compared equivalence-based probability estimation with direct probability estimation and hierarchical probability estimation, using the quality measures of Eq. 13 and Eq. 14. The numbers $h(I_1, I_0)$ and $\tilde{h}(I_1, I_0)$ were obtained by using a conditional accumulator $h(I_1, I_0)$ implemented as a 2D array, where $I_1 \in \{0, ..., M-1\}$ is the actual pixel value and $I_0 \in \{0, ..., M-1\}$ is the previous value. This accumulator, for each new value I_1 in a training set, is updated according to the following lines of pseudo-code:

$$
\begin{aligned}
I_1 &= \text{get\_next\_value}() \ ; \\
h(I_1, I_0) &= h(I_1, I_0) + 1; \\
I_0 &= I_1;
\end{aligned}
\tag{15}
$$

where, at the program start, each element of h, as well as I_0 are initialized to zero. Accumulator values $\tilde{h}(I_1, I_0)$ were obtained accordingly.

4.1 Experiment 1: Equivalence-based and direct probability estimation

Equivalence-based probability estimates were calculated according to:

$$\hat{p}_e\{I_1|I_0, \mathbf{h}(I_0)\} = \frac{1 + h(I_1, I_0)}{M + \sum_{j=0}^{M-1} h(j, I_0)} \tag{16}$$

where Eq. 1 (no prior knowledge about the meta-probability distribution) is generalized to conditional probabilities.

Experimental variance $\tilde{\sigma}_g^2(\tilde{\mathbf{h}}, \hat{\mathbf{p}}_e)$ (Eq. 13) and experimental entropy $\tilde{S}_g(\tilde{\mathbf{h}} : \hat{\mathbf{p}}_e)$ (Eq. 14), based on Eq. 16, were calculated for 320 images of various sizes, including Lenna, the Brodatz textures, a face database, and a set of high quality images. The experimental entropies are shown in Table 1) as a function of the size of the training set. The mean experimental entropy for an image class was obtained by averaging over all images in the specific class. For completeness, the quality measures were also calculated by using the direct probability estimate

$$\hat{p}_d\{I_1|I_0, \mathbf{h}(I_0)\} = \frac{h(I_1, I_0)}{\sum_{j=0}^{M-1} h(j, I_0)} \tag{17}$$

The corresponding experimental variances $\tilde{\sigma}_g^2(\tilde{\mathbf{h}}, \hat{\mathbf{p}}_d)$ are listed in Table 1. The experimental entropies $\tilde{S}_g(\tilde{\mathbf{h}}, \hat{\mathbf{p}}_d)$ are not listed, since $\log(\hat{p}_d\{I_1|I_0, \mathbf{h}(I_0)\}) = \infty$ for $h(I_1, I_0) = 0$ which means that an infinite experimental entropy results if $\tilde{h}(I_1, I_0) \neq 0$. A comparison of experimental variances for equivalence-based and direct probability estimation shows only small differences which favor equivalence-based probability estimation.

Type	Faces	Lenna	Brodatz	High quality
Size [bytes]	16384	65536	262144	4000000
Mean experimental variance $\tilde{\sigma}_g^2(\tilde{\mathbf{h}},\hat{\mathbf{p}}_e)$	0.0939	0.0295	0.0136	0.114
Mean experimental entropy $\tilde{S}_g(\tilde{\mathbf{h}}:\hat{p}_e)$	6.462	6.895	6.157	5.205
Mean experimental variance $\tilde{\sigma}_g^2(\tilde{\mathbf{h}},\hat{\mathbf{p}}_d)$	0.0923	0.0474	0.0137	0.117

Table 1: Results for equivalence-based and direct probability estimation.

4.2 Experiment 2: Hierarchical probability estimation

In a second experiment the fact was exploited that an image pixel value is the result of digitizing an analog brightness value. For this case, the results of Section 2 apply so that the probability values were estimated according to:

$$\hat{p}_h\{I_1|I_0, \mathbf{h}(I_0)\} = \Pi_{k=0}^{log_2(M)-1} \frac{(\sum_{j \in \mathcal{S}_k(I_1)} h(j, I_0)) + 1}{(\sum_{j \in \mathcal{S}_{k+1}(I_1)} h(j, I_0)) + 2} \tag{18}$$

Type	Faces	Lenna	Brodatz	High quality
Size [bytes]	16 384	65 536	262 144	4 000 000
Mean experimental variance $\tilde{\sigma}_g^2(\tilde{\mathbf{h}},\hat{\mathbf{p}_h})$	0.0518	0.0260	0.0103	0.112
Mean experimental entropy $\tilde{S}_g(\tilde{\mathbf{h}}:\hat{\mathbf{p}_h})$	4.731	6.072	5.780	5.076
$\tilde{\sigma}_g^2(\tilde{\mathbf{h}},\hat{\mathbf{p}_h})/\tilde{\sigma}^2(\tilde{\mathbf{h}},\hat{\mathbf{p}_e})$	0.55	0.88	0.75	0.98
$\tilde{S}_g(\tilde{\mathbf{h}}:\hat{\mathbf{p}_h})/\tilde{S}_g(\tilde{\mathbf{h}}:\hat{\mathbf{p}_e})$	0.73	0.88	0.94	0.98

Table 2: Results for hierarchical probability estimation.

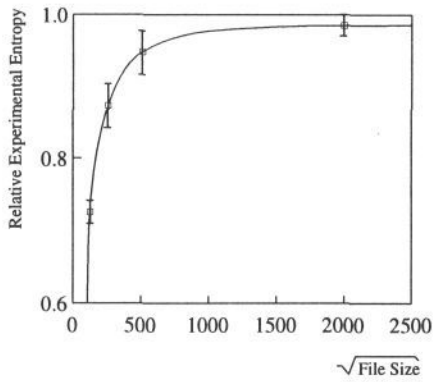

Figure 1: Relative experimental entropy $\tilde{S}_g(\tilde{h}:\hat{p_h})/\tilde{S}_g(\tilde{h}:\hat{p_e})$ against $\sqrt{\text{file size}}$

generalizing Eq. 6 to conditional probabilities.

Table 2 shows the results obtained by calculating experimental variance and experimental entropy based on hierarchical probability estimation. The last two lines compare the results in Table 1 with the results in Table 2 and show a systematic improvement for hierarchical probability estimation. This systematic improvement was found for each of the 320 test images, individually. The improvement is more prominent for small images than for large ones, as illustrated for experimental entropy in Figure 1. This is expected because, with *and* without hierarchical probability estimation, the probability estimates are less precise for small image sizes. Thus, the relative improvement is larger for smaller image sizes. Hierarchical probability estimation would, for example, lead to a 27% improvement of the average compression ratio of 128 by 128 images using an ideal lossless coder.

5 Conclusions

We predicted and experimentally verified a basic property of the joint probability distribution of digital sensor outputs. We showed that, for this class of random variables, hierarchical probability estimates are considerably better than equivalence-based or direct probability estimates.

Better probability estimates improve the results of any application which makes use of these estimates. The results of this paper, therefore, are especially important for data compression and pattern classification.

Acknowledgements

Thanks to M. J. Dürst, T. Fromherz, M. Hafner, K. Popat, and R. Sennhauser for helping to accomplish this manuscript.

References

[1] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., 1987.

[2] C. E. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal*, pp. 379-423 and pp. 623-656, **27**, 1948.

[3] M. Bichsel and A. Pentland, "Human Face Recognition and the Face Image Set's Topology", *CVGIP: Image Processing*, to be published.

[4] M. Bichsel and K. W. Ohnesorge, "The Subtleties of Probability Estimation", *University of Zurich, MultiMedia Laboratory Technical Report*, **94-09**, 1994.

[5] M. Bichsel and K. W. Ohnesorge, "The Benefits of Noise for Joint Probability Estimation", *University of Zurich, MultiMedia Laboratory Technical Report*, **94-08**, 1994.

[6] W. H. Press et al., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, 1987.

[7] G. Healey and R. Kondepudy, "CCD Camera Calibration and Noise Estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Champaign, Illinois*, 1992.

[8] "Encyclopedia of Statistical Sciences", S. Kotz and N. L. Johnson, Eds., John Wiley and Sons, **4**, 1981.

[9] Ch. Bouman and B. Liu, "Multiple Resolution Segmentation of Textured Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(2), pp. 99-113, 1991.

[10] I. H. Witten, Radford M. Neal, and John G, Cleary, "Arithmetic coding for Data Compression", *Communications of the ACM*, **30**(6), pp. 520-540, 1987.

[11] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, **77**(2), pp. 257-285, 1989.

[12] A. Bruce Carlson, *Communication Systems, An Introduction to Signals and Noise in Electrical Communication*, Third Edition, McGraw-Hill, pp. 112-113, 1987.

[13] I. N. Bronstein and K. A. Semendyayev, *Handbook of Mathematics*, 20th edition, Verlag Harri Deutsch, 1985.

[14] B. R. Frieden, *Probability, Statistical Optics, and Data Testing*, Springer,1983.

[15] R. N. Williams, *Adaptive Data Compression*, Kluwer Academic Publ., 1991.