Introduction to Part II

Human observers are an integral part of any imaging system. Image quality can be

described in purely physical terms, but optimal image quality can only be described

with reference to the performance of an imaging task. The relation between physical

image quality and diagnostic performance is the borderland between physics

and psychology known as psychophysics. The first three chapters in the perception

section deal with the current state of the art in the development and testing of psychophysical

models for predicting human observer performance. Myers provides

an overview of ideal observer models, Abbey and Buchod describe linear models,

Eckstein, Abbey and F. O. Buchod provide an overview of available models and a

roadmap for future model development. The chapter by Wilson, Jabri, and Manjeshwar

shows how human observer models can actually be used to improve the

design of dynamic fluoroscopic imaging systems. The chapter by Samei, Eyler and

Baron on the effect of anatomical backgrounds on perception provides a bridge

between the hardcore psychophysics and the softer studies of human image perception

by considering the problems of moving from statistically defined images

to real images of real people. Rolland summarizes the progress in the simulation

of realistic yet mathematically definable backgrounds for use in the models.

The methodology for assessing human performance is very important because

there is wide variability both within and between human observers even when

they perform relatively simple visibility and detection tasks. The receiver operating

characteristic (ROC) analysis has emerged as one of the major statistical analytical

tools available to characterize human performance. It can correct for variability

in the application of decision criteria. It is an active area of research in imaging.

The chapter by Metz updates this very important methodology and elaborates its

strengths and weaknesses, while the chapter by Chakraborty describes some of the

variants of classical ROC analysis and their application to special problems in diagnostic

imaging. There is also a need to go beyond the limitations imposed by the

use of the ROC analysis. One approach is to evaluate observer agreement rather

than the accuracy. The chapter by Polansky reviews the classical method for measuring

agreement in imaging and describes an alternative methodology based on

mixture distribution analysis.

Human performance even on simple tasks such as detecting tumors or fractures

is complicated by the need to locate abnormalities embedded in the complex patterns

of anatomical details on images. The chapter by Kundel deals with the role of

visual search in detection tasks. Human performance also depends upon expertisethe

combination of talent, training, and experience. Nodine and Mello-Thoms review

the status of studies of expertise in radiology and stress the implications for

training and certifying radiologists.

In the final chapter Krupinski reviews the contributions that have been made by

image perception research to the field of medical imaging.

It is the hope of the editor that these chapters mark just the beginnings of our

knowledge about the perception of medical images. They provide summaries, reviews,

and comprehensive bibliographies.What a place to start learning.

**Harold L. Kundel**

**9.1 Introduction**

The objective evaluation of imaging systems requires three important components:

(1) identification of the intended use of the resulting images, which we shall

refer to as the *task*; (2) specification of the *observer*, who will make use of the images

in order to perform the task; and (3) a thorough understanding of the statistical

properties of the objects and resulting images. With these components, a *figure of*

*merit* can be determined for evaluating the performance of the observer on the specified

task [1]. In the next sections we consider each of these elements more fully.

The *ideal observer* is a model that describes the performance of the optimum

decision maker on a given decision task. The ideal observer therefore provides an

upper bound on task performance that can be used as a gold standard in the objective

evaluation of imaging systems. Knowledge of ideal-observer performance

allows the physicist and perceptual scientist to determine when information needed

to perform a given task is readily extracted from an image by the human observer.

When ideal-observer performance is found to be far above human performance,

either the system should be redesigned to better match the human’s capabilities, or

the human observer should be augmented or even replaced by a machine reader.

In this chapter we detail the sense in which an ideal observer is optimum, describe

the calculation of the ideal-observer strategy and ensuing performance metric

for a number of tasks, and show how an imaging system can be evaluated using

figures of merit from signal-detection theory that summarize ideal-observer performance.

Results of investigations comparing human performance to that of the ideal

observer are provided for a number of visual tasks.

***9.1.1 The task***

Medical imaging tasks can be broadly categorized as either *classification* or

*estimation* tasks. In a classification task a decision is made regarding from which

class of underlying objects the data are derived. In this chapter we shall concentrate

on the binary decision task, where the image is to be classified into one of two possible

alternatives, truth state 1 (T1) or truth state 2 (T2). When the states represent

signal present (abnormal) versus signal absent (normal), the task is referred to as

signal detection. The determination of whether a lesion or tumor is present in an

image is a signal-detection task. More generally the two states are differentiated

by whatever properties of the objects in class 2 distinguish them from objects in

class 1.

An estimation task involves the quantitation of one or more parameters that

describe the object, based on the raw data. The parameter might be the size, location,

or activity of a tumor, the amount of flow in a vessel, or the cardiac ejection

fraction. In tomographic imaging the reconstruction step results in a discrete image

that is meant to estimate the spatial distribution of some characteristic of the

object, for example, the distribution of a radioactive tracer.

There is a natural relationship between classification and estimation tasks; one

can think of estimation as classification where the number of classes is the number

of possible values the parameters to be estimated can assume. Thus reconstruction

562 Ideal Observer Models of Visual Signal Detection

of a 128 × 128 image using 128 gray levels is classification into (128)3 classes!

Even so, the signal-detection theory framework for analyzing classification tasks

is generally applicable, although this example shows that the number of effective

classes represented by an estimation task can be very large.

In present times quantitation tasks typically involve a numerical algorithm applied

to an image by a computer, rather than a computation by a human. Because

the purpose of this chapter is to describe ideal-observer models and discuss their

relationship to human perception, we shall say little more about such estimation

tasks here. However, because reconstructions from tomographic data are often interpreted

by humans, we shall address ideal-observer performance on reconstructed

images to some extent below.

***9.1.2 Objects and images***

For a digital imaging system the data consist of a set of *M* discrete measurements

{*g*1*,g*2*, . . . , gM*}, where *gm* represents the *m*th measurement. Most commonly,

the data are the *M* pixels or gray levels of a digital image, although the data

might equally well be the raw (projection) data from a tomographic system. The

data values can be arranged to form a column vector **g** in an *M*-dimensional space

which we shall call data space, where the *m*th component is the value detected at

discrete element *m*. The data space can be assumed to be a Hilbert space if we

impose the usual definitions of norm and scalar product.

The data are the result of an image-formation process whereby a continuous

object *f (***r***)* is mapped to the data set **g**. (For a more detailed description of image

formation and noise in imaging systems, see Barrett andMyers [2].) This mapping

can be represented quite generally by the following expression:

**g** =***H*f**+ **n***,* (9.1)

where the imaging operator***H*** is an integral operator defined by

*gm* =

\_

*f (***r***)hm(***r***)*d2*r* + *nm, m*= 1*, . . .,M,* (9.2)

and *hm(***r***)*, called the *sensitivity function*, gives the contribution to the *m*th measurement

from the object at point **r**. The *M*-dimensional vector **n** represents the

noise in the data set.

The only assumption made in writing the imaging process as (9.2) is that the

system be linear. The sensitivity function is closely related to a matrix called the

crosstalk matrix, which describes how well particular Fourier coefficients of the

object can be recovered from a set of discrete measurements. The crosstalk matrix

is particularly useful for characterizing shift-variant imaging systems [3].

In a classification task, each truth state, often called a hypothesis, represents

a single object (in the nonrandom signal problem) or a class of objects (in the

random signal problem). The object is considered to be a continuous function *f (***r***)*

of two or three spatial dimensions, and it may have temporal dependence as well.

In writing Eq. (9.1) we regard the object as a vector **f** in a Hilbert space, say *L*2,

which we refer to as object space. The imaging operator is a mapping from object

space to data space.

The fact that the noise is represented as additive does not restrict us to additive

noise situations. It is understood that the noise is the difference between the

expected data set in the absence of noise and the actual data set. That is,

**n** = \_**g**\_ − **g***,* (9.3)

where the angle brackets denote a statistical average over all the contributions to

randomness in the data. All data are random because of measurement noise, which

might be photon noise as in the case of radiographic imaging, or thermal noise as in

magnetic resonance imaging. Additionally, the data might have some randomness

because of underlying randomness in the objects being imaged. The mean data set

under the *j* th hypothesis is then

T*j* : \_**g**\_ =***H***\_**f**\_*j* =***H*f***j ,* (9.4)

where **f***j* is the mean object in class *j* . The full probabilistic nature of the data

under truth state *j* is contained in the probability density function (PDF) on **g**, that

is, pr*(***g**|T*j )*, where the vertical bar is read “conditioned on.” (We have assumed

that the data are able to take on sufficient numbers of values so as to be modeled by

a continuous-valued vector described by a probability density function.) Another

name for pr*(***g**|T*j )* is the *likelihood* of the data given hypothesis *j* .

***9.1.3 The observer***

Statistical decision theory [4–6] invokes the concept of a decision-maker, or observer.

In the simple binary case, the decision-maker’s task is to determine which

of two classes the data set belongs to. In medical imaging the radiologist is usually

the observer of the clinical image, although many investigators are developing

computerized diagnosis systems. Generally speaking, the observer is an entity (human

or algorithm) that makes use of the data to classify it into states of truth, which

we have designated T1 and T2 in the binary decision case.

We assume that the observer’s decision rule involves no guessing or randomness

(the same data set always leads to the same decision) and no equivocation

(every data vector leads to one decision, either class 1 or class 2). It then follows

that the observer forms a scalar *decision variable*, or *test statistic*, which we shall

call *t (***g***)*, in order to classify the data. In general the formulation of the test statistic

involves nonlinear operations on the data set. We shall discuss the exact form of

the dependence of *t (***g***)* on the data in subsequent sections.

Once the test statistic is determined, the observer compares it to a *threshold \_*c

to decide between the two hypotheses. The entire set of operations—from object

to data set to decision—envisioned by the statistical decision-theory model for a

classification task is represented schematically in Figure 9.1.

***9.1.4 Decision outcomes and ROC curves***

The test statistic *t (***g***)* is itself a random variable, because it is a functional of

the multivariate random-data vector **g**. The probability density function on *t (***g***)* depends

on the state of truth, and is denoted by pr*(t (***g***)*|T*j )* for state T*j* . It is the fact

that the density functions on *t (***g***)* for each state overlap, as shown in Figure 9.2, that

renders the decision-making process interesting. Defined in terms of classification

task performance, image quality is determined by the degree of separation/overlap

of these two density functions.

Figure 9.2 shows that there are four possible decision outcomes for any value

of *tc*:

1. True positive (TP): T2 is true; observer decides T2 is true.

2. False positive (FP): T1 is true; observer decides T2 is true.

3. False negative (FN): T2 is true; observer decides T1 is true.

4. True negative (TN): T1 is true; observer decides T1 is true.

Two of the above alternatives result in the observer correctly determining the underlying

hypothesis, but we also see that two types of errors can be made. If the

problem is to decide whether signal is present or absent, and the observer says a

signal is present when it fact it is not, a type I error is made. In radar terminology

**Figure 9.1:** The entire set of operations—from object to data set to decision—envisioned

by the statistical decision theory model for a classification task.

**Figure 9.2:** The decision theory paradigm.

**Figure 9.3:** Example ROC curve showing 3 operating points: (A) conservative threshold;

(B) moderate threshold; (C) lax threshold.

this is called a *false alarm*, while in the medical literature it is called a *false positive*.

When the signal is present but the observer chooses the noise-only alternative,

we say a *miss*, or *false negative* has occurred. This is known as a Type II error. It

follows that the *true-positive fraction* (TPF) is the probability of a true-positive decision.

This is known as the *sensitivity* in the medical imaging literature. The *falsepositive*

*fraction* (FPF) is the probability of deciding in favor of the signal-present

hypothesis when the signal is not there. In medical applications the *specificity* is

often reported, which is given by (1-FPF).

The observer’s threshold *λc* specifies the *operating point* of the observer, that is,

the (TPF, FPF) pair. By varying the threshold, a family of (TPF, FPF) points can be

generated.A graph of these points is known as the *receiver operating characteristic*

curve, or ROC curve, an example of which is given in Figure 9.3. Metz discusses

the properties and measurement of ROC curves in Chapter 15 of this volume [7].