Plausible Reasoning and the Theory of Evidence

George Reynolds, Deborah Strahman

Nancy Lehrer, Les Kitchen

COINS Technical Report 86-11

April 1986

ABSTRACT

In this paper we describe the mathematical foundations of a knowledge representation and evidence combination framework and relate it to the theory of evidential reasoning as developed by Dempster and Shafer. Although our discussion takes place in the context of computer vision, the results are applicable to problems in knowledge representation and data interpretation. Our representation, called pl-functions, and a simple multiplicative combination rule is shown to be equivalent to a sub-class of the family of mass-functions as described by Shafer with Dempster's rule as the combination function. However the simpler combination rule has a complexity which is linear with respect to the number of elements in the frame of discernment. This is a tremendous computational advantage over the general theory which provides a combination rule exponential with respect to the number of objects over which we are reasoning. We also discuss a method which allows our representation to be automatically generated from statistical data.

Topics: Perception and Signal Understanding, Theory of Evidence, Dempster's Rule, Possibilistic Reasoning, Belief Functions.

1. Introduction

Many problems in AI have associated with them the problems of how to build a structural hierarchy of complex objects and how to combine information from multiple knowledge sources. Specifically, in computer vision these problems take the form of how to combine information derivable from various features of primitive image events or tokens (lines, curves, regions and surfaces) in order to make inferences about more complex image events (objects in the domain). Much recent work has addressed the problems of how to convert information about these primitives into "evidence" for higher-level events and to invoke some more abstract "evidence combination" rule in order to fuse the data provided by various knowledge sources. These include Shafer's [1976b] theory of evidence (see also Lowrance and Garvey [1983]), voting schemes (Hanson et.al. [1984]), Bayesian methods (Duda and Hart [1973]), constraint propogation techniques (Kitchen and Rosenfeld [1984]), maximum entropy techniques (Geman and Geman [1984], Shastri and Feldman [1984]), fuzzy sets (Zadeh [1978]), and various ad hoc heuristic methods (Shortliffe [1976]).

In this paper we will examine a method for representing and combining evidence and relate it to the theory of evidence as proposed by Shafer. Before motivating the mathematical formalism we will discuss the kind of evidence combination process to which we see our formalism being applied.

The kind of reasoning that an image understanding system is asked to perform is often not unlike that of a detective given the task explaining the circumstances surrounding a crime. The detective is presented with some raw data and asked to make inferences about who committed the crime. If the crime was a murder, the detective could invoke three different kinds of reasoning methodologies.

First, some observations would allow significant reductions of the search space (all people who might have committed the murder), or at least suggest parts of the search space which are more likely, without yielding any specific information about individuals within that subset actually did it. A neighbor says that she has heard

members of the family in heated arguments in the last few weeks suggesting that some member of the family actually committed murder. Or a woman's glove is found on the stairs leading away form the site of the crime suggesting that a woman committed the crime. However there is no information in this kind of evidence as to which woman in the family did it nor a likelihood associated with any particular suspect.

Second, after aquiring some initial evidence the detective may make a hypothesis that Sue committed the crime and spend some time looking for confirming or disconfirming evidence for this hypothesis. The detective may try to obtain evidence that Sue has or was seen wearing gloves matching those found at the location of the murder.

Finally the detective will be required to perform some reasoning at a level above the previous two kinds and look for groupings of evidence each of which is consistent with every other element of the group. For example the detective may discover that there is evidence that every woman in the family was out of town on the day of the crime, suggesting that either this evidence or the glove evidence, but not both, are parts of a consistent group of evidence which points to the culprit.

A computer vision system often needs to performs analogous kinds of reasoning. Bottom up processes generate tokens whose features are consistent with many possible object hypotheses. Whatever the evidence, the system needs to make specific hypotheses, explore their consequences, check for consistency and explain any inconsistencies. Many objects will share features and relations within their descriptions in the knowledge base, and the unique identification of an object will arise only by combining the "evidence" from many feature spaces that these tokens provide for these descriptions and relations.

In this paper we will develop a representation which supports all three types of inference described above. Our discussion has three components. The first part is a method for representing evidence using what will be called pl-functions ("pl" for plausibility). These functions associate measurable properties (e.g. features) of the image events, via knowledge sources, to labels which are to be assigned to abstractions of these image events. pl-functions capture the extent to which a label

is plausible given some feature measurement. We will also briefly discuss a proposal for automatically generating this representation when certain types of statistics are available.

In the second part of the discussion we will define a function

$$MA\text{: pl-functions}(\Theta) \rightarrow \text{mass-functions}(\Theta)$$

where Θ is a frame of discernment (see section 2 for all the relevant definitions), pl-functions(Θ) is all pl-functions on Θ and mass-functions(Θ) is all mass functions on Θ. We will show that this function has the property that the combination of pl-functions by multiplication is equivalent to combination of mass functions by Dempster's rule. Observe that the simple multiplicative combination process is **linear** with respect to the number of objects about which we are reasoning. This is a tremendous computational advantage over the general theory which provides a combination rule whose complexity is exponential with respect to the number of elements in the frame of discernment.

Our construction is related to a method of evidential inference termed "conditional embedding" by Shafer [1982]. For cases where the pl-function is derived from likelihoods formed from independent frequency distributions this function can be shown to be equivalent to the specific case of conditional embedding developed by Smets [1978].

Thus we will be describing a way of associating a feature measurements with a mass function. Methods for making such an association have been made before (Lowrance [1982], Wesley and Hanson [1985]). However these methods require that the range of values over which the mass functions are defined be either explicitly or implicitly discretized into "feature propositions" or subintervals of the feature variable such as "low", "medium" and "high" feature values. In our approach no such artifical discretisation is required. The mapping from feature value to mass function is defined in terms of feature values of arbitrarily fine quantisation and can readily be extended to the continuous case.

In the last part of our discussion we will examine the relationship between the mappings

$$MA\text{: pl-functions}(\Theta) \rightarrow \text{mass-functions}(\Theta)$$

given by the construction given in section 3, and

$$PL: \text{mass-functions}(\Theta) \rightarrow \text{pl-functions}(\Theta)$$

which assigns to any mass function the pl-function given by the plausibilities on singletons. Our question is: to what extent do the plausibilities of a mass function on the singleton sets capture all the relevant information about the mass function? We will show that for any consistent mass function, (the plausibility of some singleton is 1) then the rank order given by the plausibilities is identical with the rank order given by a more complex measure defined in terms of support and plausibility.

2. Basic Definitions from the Theory of Evidence

In this section we will briefly review some of the basic definitions from the theory of evidence (Shafer [1976]). Suppose we are presented with a question and a finite set, Θ, consisting of possible answers to the question, only one of which is the correct one. Then for each $o \in \Theta$ the proposition of interest is precisely of the form *"The correct answer is o"*. A set will be called a *frame of discernment* when its elements are interpreted as possible answers to a particular question, and we know that exactly one of the answers is correct. Each subset $P \subseteq \Theta$ can be interpreted as a proposition which states : *"The correct answer is in the set P"*. Thus the set of all propositions relevant to finding the correct answer is in a one to one correspondence with the set of subsets of Θ, i.e. 2^{Θ}.

Definition: A mass function is a function

$$m: 2^{\Theta} \rightarrow [0, 1]$$

so that

$$m(\phi) \neq 0$$

$$\textit{and} \quad \sum_{A \subseteq \Theta} m(A) = 1$$

Given set $P \subseteq \Theta$, $m(P)$ should be interpreted as the amount of belief or evidence M has that P is the set every element of which the evidence supports as being the the correct answer. Below we will see that each measurement we make on an environment will generate a mass function which is the evidence that that measurement provides as to which sets contain the correct answer. Thus measurements of two different features will provide two different mass functions each of which is the distribution of a unit of belief over sets which contain the correct answer. Dempster's rule is a way of combining mass functions, the result being another mass function which focuses

the mass on the set which **both** measurements support as the set containing the possibly correct answers.

Dempster's Rule: If m_1 and m_2 are mass functions then

$$m_1 \oplus m_2(C) = \frac{\sum_{A \cap B = C} m_1(A) \cdot m_2(B)}{1 - k}$$

where

$$k = \sum_{A \cap B = \phi} m_1(A) \cdot m_2(B).$$

$m_1 \oplus m_2$ is called the **combination** of m_1 and m_2, k is called the **conflict value,** and the combination is defined if and only if $k \neq 1$. If we interpret the two mass functions as each distributing a unit of belief for the sets which a knowledge source believes contain the correct answer, then $k = 1$ if and only if the evidence provided by m_1 flatly contradicts the evidence provided by m_2. In general the conflict value is a measure of the extent to which two bodies of evidence contradict each other with $k = 0$ precisely when they are consistent.

The conflict value k may be non-zero for one of two reasons. On the one hand the evidence represented by the mass function may be in error, or on the other, one of the assumptions implied by the representation of knowledge within the frame of discernment has been violated. In the latter case it may be that the frame of discernment and the process of generating the mass functions needs to be modified to correctly reflect the assumptions of the domain, or that an event has occurred which was not included in the frame and Θ needs to be enlarged.

If a set $A \subseteq \Theta$ is assigned mass t, then any set B with $A \subseteq B \subseteq \Theta$ should believe to an amount at least t that it too contains the right answer. In addition any set $C \subseteq \Theta$ with $C \cap A = \phi$ should have the extent to which the evidence refutes C as containing the right answer reduced by at least t. This leads to the following definition:

Definition: Given a mass function $m: 2^\Theta \rightarrow [0, 1]$, the **support** and **plausibility** of

each $A \epsilon 2^{\Theta}$ are defined respectively as follows:

$$spt(A) = \sum_{X \subseteq A} m(X)$$

$$pls(A) = 1 - \sum_{X \cap A = \phi} m(X).$$

In summary, the $spt(A)$ is the total positive impact of the evidence on A, and the $pls(A)$ is the extent to which the evidence fails to refute A. It is always the case that $0 \leq spt(A) \leq pls(A) \leq 1$.

Definition: Given a mass function $m: 2^{\Theta} \rightarrow [0, 1]$, the **decisiveness** of each $A \epsilon 2^{\Theta}$ is defined as

$$dec(A) = spt(A) - (1 - pls(A)).$$

See Wesley and Hanson [1985] for a more complete description of these and related measures. Note that $dec(A)$ is a number between -1 and 1 and it has the following interpretation. If the $spt(A)$ is close to 1 then $dec(A)$ is close to 1 and of the $pls(A)$ is close to 0 then $dec(A)$ is close to -1. Thus if $dec(A)$ is close to 1 then the evidence supports A; if $dec(A)$ is close to -1, the evidence tends to refute A, and if $dec(A)$ is close to 0 then the evidence is indecisive.

One simple decision criterion is to compute the decisiveness for each singleton of Θ and take the element which has the maximum value. That is select $a \in \Theta$ where

$$dec(\{a\}) = max\{dec(\{x\}) \mid x \in \Theta\}.$$

See Wesley and Hanson [1985] for a more complete description of these and related measures.

3. Converting Measurements into Mass Functions

In many image understanding problems we are often faced with the problem of labeling some segmentation of an image given some statistical information from various feature spaces FS, and their relationship to the labels Θ. In particular we may have information about the relative frequencies of various features with respect to various labels, ie. we have some information about the distribution $p(a \mid f)$ where $f \in FS$ and $a \in \Theta$. Even if the statistics obtained are inaccurate, there is still a significant amount of knowledge contained in these distributions. If the frequency is high, for some value f, then at least we don't want to rule out the possibility that the correct label to be assigned is a. In addition the feature value f may occur frequently for many objects and so the knowledge we have is of the form: *Given an observation f, then we don't want to rule out the possibility that the correct label is in the set A of labels for which that feature occurs frequently.* Combining the information from many such knowledge sources then can significantly reduce the search space of plausible labels.

What is needed for each $a \in \Theta$ is a function $FS \rightarrow [0,1]$ which defines how plausible the label a is given some feature $f \in FS$. In this section we will show how such a function, which we call a pl-function, yields a mass function on Θ and in the next section we will show how pl-functions can be derived from statistical data.

Definition: Given a frame of discernment Θ and a feature space FS, a pl-function

$$pl(a \mid f) : FS \rightarrow [0,1]$$

is a function defined on a feature space FS for each $a \in \Theta$ which has the interpretation: $pl(a \mid f)$ is the extent to which we don't want to rule out a if we make the observation $f \in FS$.

Definition: A knowledge source is a function

$$ks : FS \rightarrow M(2^{\Theta})$$

where $M(2^{\Theta})$ is the set of all mass functions on Θ.

Definition: A context is a specification of a set Θ and a collection of knowledge sources

$$ks_1 : FS_1 \rightarrow M(2^{\Theta}), \ldots, ks_n : FS_n \rightarrow M(2^{\Theta}).$$

A frame of discernment is designed to capture the relationships between the objects in some context and the features in that context which pertain to reasoning about those objects. As the context changes, the objects, the features and the relationships between the features and the objects can be expected to change.

Each feature space can be thought of as containing quantities that are associated with some observable and quantifiable aspect of the knowledge we are bringing to bear on the problem of answering the question which the context is designed to answer. The set of all feature spaces of potential interest and their knowledge sources forms a context. In general this includes any aspect of a domain or world about which information may be obtained in order to help decide which answer is correct. In our approach to reasoning about one's environment, various types of knowledge sources provide the partially processed information, based on their environmental observations, about the "evidence for" or "belief in" the propositions represented by the subsets of Θ.

Suppose we are given a set Θ and for each $a \in \Theta$ a pl-function. We will now give a construction which generates a knowledge source from this pl-function. For each $A \subseteq \Theta$ define

$$m_0(A \mid f) = \prod_{a \in A} pl(a \mid f) \prod_{a \in \Theta - A} (1 - pl(a \mid f)).$$

Now the function $m_0(A \mid f)$ is **not** a mass function since

$$m_0(\phi \mid f) = \prod_{a \in \Theta} (1 - pl(a \mid f))$$

is not necessarily equal to zero. However the function does have the other two properties of a mass function:

1. $0 \leq m_0(A \mid f) \leq 1$,

2. $\sum_{A \subseteq \Theta} m_0(A \mid f) = 1$.

The first statement is obvious and the following lemma directly implies the second.

Lemma 3.1 *If $(x_1, ..., x_n)$ is a sequence of numbers and $N = \{1, ..., n\}$ then*

$$\sum_{A \subseteq N} \prod_{i \in A} x_i \prod_{j \in N-A} (1 - x_j) = 1$$

where we define

$$\prod_{\phi} x_i = 1$$

Proof: The proof is by induction. Observe that if $N = \{1\}$ then

$$\sum_{A \subseteq N} \prod_{i \in A} x_i \prod_{j \in N-A} (1 - x_j) = x_1 + 1 - x_1 = 1.$$

The nature of the induction is clarified by considering $N = \{1, 2\}$. In this case

$$\sum_{A \subseteq N} \prod_{i \in A} x_i \prod_{j \in N-A} (1 - x_j) = (1 - x_1)(1 - x_2) + x_1(1 - x_2) + x_2(1 - x_1) + x_1 x_2$$

$$= (1 - x_1 + x_1)(1 - x_2) + (1 - x_1 + x_1)x_2 = 1.$$

In general

$$\sum_{A \subseteq (N+1)} \prod_{i \in A} x_i \prod_{j \in (N+1)-A} (1 - x_j)$$

$$= \left(\sum_{A \subseteq N} \prod_{i \in A} x_i \prod_{j \in N+1-A} (1 - x_j) \right) (1 - x_{n+1}) + \left(\sum_{n+1 \in A \subseteq N+1} \prod_{i \in A} x_i \prod_{j \in N+1-A} (1 - x_j) \right) x_{n+1}.$$

From the induction hypothesis this expression equals

$$1 - x_{n+1} + x_{n+1} = 1.$$

This completes the proof.

Consider what it means for the empty set to receive a non-zero value in terms of the pl-functions generating m_0. It means simply that the consensus of opinion of the pl-functions is that to some degree the feature value in question rules out every element of Θ with respect to the current state of the knowledge base (as represented by the pl-functions). This could be either because the knowledge source is in error or the knowledge base is incomplete. The decision as to which of these conditions holds is external to the processes of the inference network. All that should be required of it is that it return (partially) the answer *unknown*.

Therefore we add to Θ a new element unk and define

$$m(A \cup \{\text{unk}\} \mid f) = m_0(A \mid f).$$

This then, is a mass function on $\Theta \cup \{\text{unk}\}$

We could have recast our definition by defining a new object (a pre-mass function?) which is allowed to assign non-zero mass to the empty set (see Hummel [1985]). Dempster's rule can be defined for these objects (just take out the re-normalization) and there is a simple mapping between these objects and mass functions. However, this approach requires doubling the notation. The addition of {unk} requires no change in notation or conceptualisation, eliminates the need to re-normalise until it is appropriate and the "conflict" value generated by Dempster's rule is simply the mass assigned to {unk}.

The following theorem summarizes some of the relationships between the pl-functions and the generated mass function.

Theorem 3.1 *Suppose we are given a set $\Theta \cup \{unk\}$ and for each $a \in \Theta$ a pl-function*

$$pl(a \mid f) : FS \to [0, 1],$$

from some feature space FS to $[0, 1]$. Then defining the mass function $m(A \cup \{unk\} \mid f)$ as above,

1. *$spt(A) = 0$ if $unk \notin A$,*

2. $pls(\{a\}) = pl(a \mid f)$ *for any* $a \in \Theta$,

3. $spt(\{a, unk\} = \Pi_{s \in \Theta}(1 - pl(x \mid f)) + pl(a \mid f) \Pi_{s \neq a}(1 - pl(x \mid f))$ *for any* $a \in \Theta$,

4. $pls(A \cup \{unk\}) = 1$ *for any* $A \subseteq \Theta$.

Proof: 1 is clear since if $unk \notin A$ then A is assigned zero mass by definition.

For any $a \in \Theta$

$$pls(\{a\}) = \sum_{a \in A} m(A \mid f) = \sum_{a \in A} \prod_{b \in A} ps(f \mid b) \prod_{b \in \Theta - A} (1 - ps(f \mid b)).$$

$$= ps(f \mid a) \sum_{B \subseteq \Theta - \{a\}} \prod_{b \in B} ps(f \mid b) \prod_{b \in (\Theta - \{a\}) - B} (1 - ps(f \mid b))$$

and according to the Lemma this is equal to $ps(f \mid a)$. This completes the proof of 2.

For the proof of 3, observe that

$$spt(\{a, unk\} = m(\{unk\}) + m(\{a, unk\})$$

which by definition is the expression given.

Finally for 4, note that $unk \in A$ for every set receiving non-zero mass. Thus $pls(A \cup \{unk\}) = 1$. This completes the proof.

Given a mass function as defined above, we can define a function on 2^Θ by the formula

$$m\text{-}norm(A \mid f) = \frac{m(A \cup \{\text{unk}\} \mid f)}{1 - m(\{\text{unk}\} \mid f)}$$

for any non-empty set A and $m\text{-}norm(\phi \mid f) = 0$.

Theorem 3.2 *m-norm is a mass function on* 2^Θ.

Proof: Simply observe that

$$\sum_{A \subseteq \Theta} \frac{m(A \cup \{unk\})}{1 - m(\{unk\})} = \frac{1}{1 - m(\{unk\})}.$$

Thus

$$\sum_{A \neq \phi} m\text{-}norm(A \mid f) = \frac{1}{1 - m(\{unk\})} - \frac{m(\{unk\})}{1 - m(\{unk\})} = 1.$$

Thus we have defined a function

$$MA\text{: pl-functions}(\Theta) \rightarrow \text{mass-functions}(\Theta)$$

which assigns to every pl-function *pl* the mass-function *m-norm*. In the section 5 we will discuss some of the properties of this function and its relationship to Dempster's rule.

Below we present an example of the process of converting pl-values into a mass function. In this example $\Theta = \{a, b, c\}$. We have displayed the mass function, the renormalized mass function, together with the support, plausibility and decisiveness for that mass function.

```
pl-values
=========

((a . 0.9) (b . 0.5) (c . 0.1))

Mass function
=============
(unk a b)       0.4050
(unk a)         0.4050
(unk b)         0.0450
(unk)           0.0450
(unk a c)       0.0450
(unk a b c)     0.0450
(unk c)         0.0050
(unk b c)       0.0050
```

Renormalized mass function

unknown value = 0.045

(a b)	0.4241
(a)	0.4241
(b)	0.0471
(a c)	0.0471
(a b c)	0.0471
(c)	0.0052
(b c)	0.0052

Spt, Pls and Dec for renormalized mass functions

subset	spt	pls	dec
(a b c)	[1.000,	1.000]	1.000
(b c)	[0.058,	0.576]	-0.366
(a c)	[0.476,	0.953]	0.429
(a b)	[0.895,	0.995]	0.890
(c)	[0.005,	0.105]	-0.890
(b)	[0.047,	0.524]	-0.429
(a)	[0.424,	0.942]	0.366

4. Generating Pl-functions from Statistical Data

In this section we describe a way of generating pl-functions from statistical data. Our goal is to describe the areas in feature space where the feature values for a given object tend to cluster. In those areas we want the pl-values to be close to 1. Moreover we want this function to be insensitive to the sample sizes used to construct the feature distributions.

Consider the ratio of the number of instances of object a with feature value f (denoted $h(f \wedge a)$) to the number of instances of any object with feature value f (denoted $h(f)$). This is by definition an estimate of the conditional probability $p(a \mid f)$,

$$\hat{p}(a \mid f) = \frac{h(f \wedge a)}{h(f)}$$

where we use $\hat{p}$ since we are dealing only with estimates of the true probabilities. Similarly defined is the relative frequency of seeing a (the "percentage" of a occuring in the sample),

$$\hat{p}(a) = \frac{\sum_f h(f \wedge a)}{\sum_f h(f)}$$

and the relative frequency of seeing the feature value f

$$\hat{p}(f) = \frac{h(f)}{\sum_f h(f)}.$$

As mentioned earlier, one desirable characteristic of a pl-function is that it be relatively invariant with respect to the size of the sample set used to model the feature distribution. That is, two distributions differing only in the size of the sample set over which they are defined, should have the same pl-function. One way of obtaining this behavior is to use $\hat{p}(a)$ as a decision threshold. If the value of $\hat{p}(a \mid f)$ is at least as large as the estimate of seeing a, $\hat{p}(a)$, we want the pl-value to be 1. Intuitively, if

we have more reason to believe in the occurence of a after the observation f than we did before the observation f, then we do not want to rule out a as being the reason f was observed. (In fact there may be factors other than $\hat{p}(a)$ which might be used in order to make the pl-function more or less conservative.)

This suggests the following definition of a pl-function,

$$pl(a \mid f) = min(1.0, \frac{\hat{p}(a \mid f)}{\hat{p}(a)}).$$

Now consider the behavior of

$$\frac{\hat{p}(a \mid f)}{\hat{p}(a)}$$

with respect to two objects with the same distribution but with different sample sizes. Let $\hat{p}(f \wedge a_2) = \alpha\hat{p}(f \wedge a_1)$ and $\hat{p}(a_2) = \alpha\hat{p}(a_1)$ where α is some constant. Then

$$\frac{\hat{p}(f \wedge a_2)}{\hat{p}(f)\hat{p}(a_2)} = \frac{\alpha\hat{p}(f \wedge a_1)}{\hat{p}(f)\alpha\hat{p}(a_1)} = \frac{\hat{p}(f \wedge a_1)}{\hat{p}(f)\hat{p}(a_1)}$$

which implies that the function $pl(a \mid f)$ is independent of the sample size. See the top two distributions in Figure 4.1.

Suppose now that $h(f \wedge a_i)$, $i = 1, ..., n$ are Gaussian's with means μ_i and variances σ_i. The next theorem states that if we are "close enough" to the mean μ_i and "far enough" away from the other means μ_j then the plausibility of a_i is equal to 1.

Theorem 4.1 *Suppose $h(f \wedge a_i)$, $i = 1, ..., n$ are Gaussians with means μ_i and standard deviations σ_i. Then if*

$$|f - \mu_i| \leq \sqrt{2}\sigma_i \text{ and } |f - \mu_j| \geq \sqrt{2}\sigma_i, \; j \neq i,$$

$$pl(a_i|f) = 1.$$

In other words, if the feature value f is within $\sqrt{2}\sigma_i$ of μ_i and farther than $\sqrt{2}\sigma_i$ from all other μ_j, then the pl-value of a_i (its "plausibility") is equal to 1, and this is

independent of the sample sizes of the distributions. Of course the choice of $\hat{p}(a)$ is still somewhat arbitrary and other choices for this value can make the pl-function more or less conservative. See figure 4.1 where the behavior of the function is illustrated by the dotted lines if $\hat{p}(a)^2$ is chosen.

Proof: We are given that

$$h(x \wedge a_i) = k_i e^{-(x-\mu_i)/2\sigma_i^2}, \ i = 1, ..., n$$

and

$$h(x) = \sum_i h(x \wedge a_i).$$

Note that

$$\int_{\infty}^{\infty} h(x \wedge a_i) dx = \sqrt{2\pi} k_i \sigma_i,$$

so it suffices to show that

$$\frac{e^{-(x-\mu_i)/2\sigma_i^2}}{\sum_j k_j e^{-(x-\mu_j)/2\sigma_j^2}} \geq \frac{\sigma_i}{\sum_j k_j \sigma_j}.$$

Using our assumptions, this boils down to showing that

$$\frac{e^{-1}}{\sum_j k_j \sigma_i^2/\sigma_j^2} = \frac{1}{\sum_j e^{1-\sigma_i^2/\sigma_j^2}} \geq \frac{\sigma_i}{\sum_j k_j \sigma_j} = \frac{1}{\sum_j k_i \frac{\sigma_j}{\sigma_i}}$$

The inequality is equivalenet to showing that for each j

$$e^{1-\sigma_i^2/\sigma_j^2} \leq \sigma_j/\sigma_i,$$

and this follows from the inequality

$$\sigma_i^2/\sigma_j^2 - 1 \geq \ln(\sigma_i/\sigma_j).$$

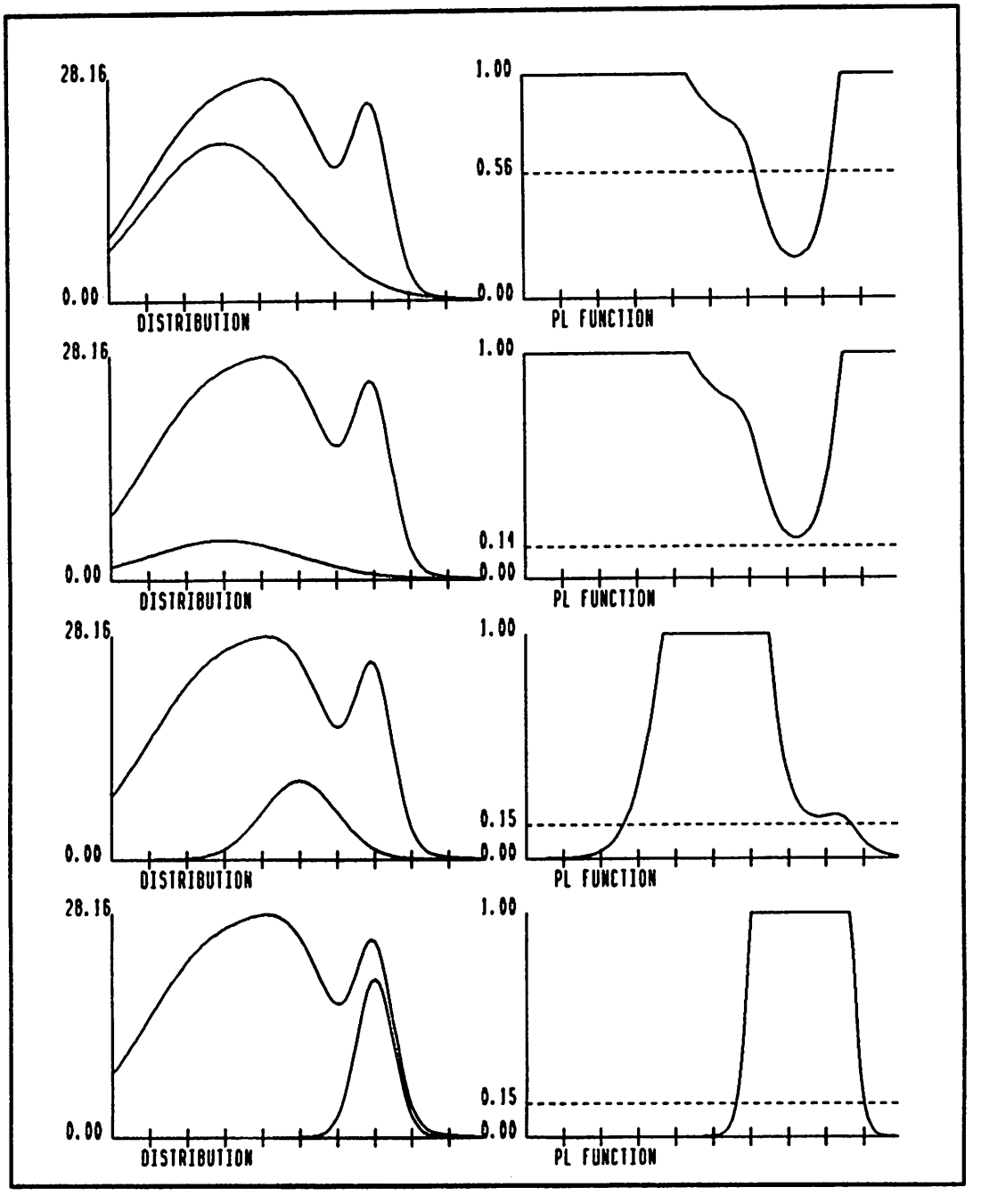

Figure 4.1: On the left of this figure are displayed a distribution $h(f)$ which is sum of four Gaussian distributions representing $h(f \wedge a)$ shown separately from top to bottom. On the right, the four pl-functions $pl(a|f)$ are displayed. Note that the top two distributions have the same mean and variance, and their pl-functions are identical. The dotted horizontal lines pass through $\hat{p}(a)$ and the area below them describes how the pl-functions would appear if $\hat{p}(a)^2$ was used in the definition of pl instead of $\hat{p}(a)$.

5. Pl-functions and Dempster's Rule

In this section we make the connection between pl-functions and Dempster's rule. In particular we will show that combining pl-functions by term-wise product and generating a mass function yields the same result as individually generating mass functions and combining using Dempster's rule.

It is useful to consider this result in a very simple case. Consider a mass function $M: 2^\Theta \to [0,1]$ for which there exists some $A \subseteq \Theta$ with $m(A) = 1$. For this paragraph call such mass functions "absolute". Let $\chi_A: \Theta \to [0,1]$ be the characteristic function of $A \subseteq \Theta$. If we view the characteristic function as a pl-function on Θ and generate the mass function χ_A-*norm*, then this mass functions is absolute and its only non-zero value is on A. It is now easy to see that when $A \cap B \neq \phi$

$$\chi_A\text{-}norm \cdot \chi_B\text{-}norm = \chi_{A \cap B}\text{-}norm,$$

in other words, multiplication of characteristic functions and generating a mass function yields the same result as first generating the mass function and applying Dempster's rule. Our result is that this is true for arbitrary pl-functions.

The first two results of this section are useful for deriving computationally efficient ways of computing Dempster's rule, supports and plausibilities. The first observation is that if we are presented with a set of mass functions on $\Theta \cup \{\text{unk}\}$ to combine, and they only assign non-zero mass to subsets of $\Theta \cup \{\text{unk}\}$ which contain unk, then the amount of mass which accumulates on $\{\text{unk}\}$ is exactly the conflict value of the n-wise combination as defined by Shafer (see also Hummel [1985]). Next we observe that normalization and combination commute with each other.

Theorem 5.1 *Suppose m_1 and m_2 are mass functions derived from the pl-functions $pl_1(a \mid f)$ and $pl_2(a \mid f)$. Then*

$$(m_1 \oplus m_2)\text{-}norm = m_1\text{-}norm \oplus m_2\text{-}norm$$

Proof: Let m_Θ be the mass function on $\Theta \cup \{unk\}$ be defined by $m_\Theta(\Theta) = 1$ and 0 otherwise. then $m\text{-}norm = m \oplus m_\Theta$, in other words, $m\text{-}norm$ is obtained by conditioning on Θ. Thus

$$m_1\text{-}norm \oplus m_2\text{-}norm = m_1 \oplus m_\Theta \oplus m_2 \oplus m_\Theta.$$

Using the associativity and commutativity of Dempster's rule we obtain

$$m_1\text{-}norm \oplus m_2\text{-}norm = m_1 \oplus m_2 \oplus m_\Theta = (m_1 \oplus m_2)\text{-}norm.$$

This completes the proof.

The next theorem shows that with respect to the elements of Θ the supports and plausibilities on the singletons can be computed directly from the pl-functions without the need of the power set. Thus any decision rule based on the support and plausibility has a complexity proportional to the number of elements of Θ (see Wesley and Hanson [1985]).

Theorem 5.2 *Suppose $m(A \cup \{unk\})$ is a mass functions derived from the pl-function pl. Then with respect to the mass function m-norm,*

$$pls(\{a\}) = \frac{pl(a \mid f)}{1 - \prod_{a \in \Theta} (1 - pl(a \mid f))},$$

$$spt(\{a\}) = \frac{pl(a \mid f) \prod_{b \in (\Theta - \{a\})} (1 - pl(b \mid f))}{1 - \prod_{a \in \Theta} (1 - pl(a \mid f))}.$$

Proof: Both equations follow directly from 2. and 3. of theorem 3.1 and the definition of *m-norm*.

The next two theorems show that term-wise product of pl-functions is equivalent to combining using Dempster's rule. We first observe that every mass function generated by a pl-function is a separable mass function.

Theorem 5.3 *Suppose we are given a set Θ and for each $a \in \Theta$ a pl-function*

$$pl(a \mid f) : FS \to [0,1],$$

from some feature space FS to $[0,1]$. Then the mass function m-norm generated by this pl-function is a separable mass function.

The proof of this result is contained in the proof of the following Lemma.

Lemma 5.1 *Suppose $(x_1, ..., x_n)$ is a sequence of numbers with $0 \leq x_i \leq 1$ and $N = \{1, ..., n\}$. For each i, let*

$$x_i^j = \begin{cases} x_i & \text{if } i = j \\ 1 & \text{otherwise.} \end{cases} \tag{5.1}$$

Define

$$m_0(A) = \prod_{i \in A} x_i \prod_{j \in N - A} (1 - x_j)$$

and

$$m_0^i(A) = \prod_{k \in A} x_k^i \prod_{l \in N - A} (1 - x_l^i).$$

Then

$$m_0(A) = \sum_{A = \bigcap A_i} \prod m_0^i(A_i).$$

A few words are in order concerning this lemma. First note that m_0 is the function we defined above to generate mass functions from possibility functions. Second, observe that any simple mass function can be generated from a sequence of numbers $(x_1, ..., x_n)$ where at most one of the numbers is not equal to 1, by the formula defining m_0. Indeed suppose $(x_1, ..., x_n)$ is a sequence of numbers with $x_i \neq 1$ and $x_j = 1$ for all other j. Then defining m by the formula

$$m(A) = \prod_{i \in A} x_i \prod_{j \in N - A} (1 - x_j)$$

then it follows that

$$m(A) = \begin{cases} x_i & \text{if } A = N \\ 1 - x_i & \text{if } A = N - \{i\} \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Now for the proof of Lemma 5.1. First note that in the expression

$$m_0(A) = \sum_{A = \bigcap A_i} \prod m_0^i(A_i)$$

the only expressions $m_0^i(A_i)$ which are non-zero arise from sets A_i where $A_i = N$ or $A_i = N - \{i\}$ (see the preceding paragraph). Indeed if we form the sets $A_i = N - \{i\}$ then

$$A = \bigcap_{i \in N - A} A_i,$$

and moreover $m_0^i(N - \{j\}) \neq 0$ only when $i = j$. Thus $m_0(A)$ involves only one summand which reduces to

$$\prod_{i \in A} m_0^i(N) \prod_{j \notin A} m_0^j(N - \{j\})$$

which equals

$$\prod_{i \in A} x_i \prod_{j \notin A} (1 - x_j)$$

as desired. This completes the proof.

Finally the main result of this section. Its proof briefly rests on theorem 5.3, in particular that a mass function generated using our rule can be derived by combining mass functions whose only focal element other than Θ is the complement of a singleton.

Theorem 5.4 *Suppose we are given a set Θ and for each $a \in \Theta$ pl-functions*

$$pl_1(a \mid f) : FS \to [0, 1],$$

and

$$pl_2(a \mid f) : FS \to [0, 1],$$

from some feature space FS to $[0,1]$. Let

$$pl_3(a \mid f) = pl_1(a \mid f) \cdot pl_2(a \mid f)$$

Then defining the mass functions m_1, m_2 and m_3 in terms of these pl-functions as above,

$$m_3 = m_1 \oplus m_2.$$

Proof: First consider the mass functions m_1', m_2' and m_3' generated by the possibility functions where every element is 1, except the i-th element, which is $ps_1(f \mid a)$, $ps_2(f \mid a)$ and $ps_1(f \mid a) \cdot ps_2(f \mid a)$ respectively. Let's abbreviate these values a_1, a_2 and a_3. Then

$$m_1'(\Theta) = a_1,\ m_1'(\Theta - \{a\}) = 1 - a_1,$$

$$m_2'(\Theta) = a_2,\ m_2'(\Theta - \{a\}) = 1 - a_2$$

$$m_3'(\Theta) = a_1 a_2,\ m_3'(\Theta - \{a\}) = 1 - a_1 a_2.$$

On the other hand

$$m_1' \oplus m_2'(\Theta - \{a\}) = (1 - a_1)(1 - a_2) + a_1(1 - a_2) + a_2(1 - a_1),$$

and a simple calculation shows that this is equal to $1 - a_1 a_2$. The proof is now completed by expanding m_1 and m_2 into their simple components (using Lemma 2) and using the associativity and commutativity of Dempster's rule.

6. Consistent Mass Functions and the Completness of Pl-functions

In this section we will briefly consider the relationship between the mapping

$$MA\text{: pl-functions}(\Theta) \rightarrow \text{mass-functions}(\Theta)$$

and the reverse mapping

$$PL\text{: mass-functions}(\Theta) \rightarrow \text{pl-functions}(\Theta)$$

where PL is defined by assigning to a mass function the function which assigns the plausibility on each singleton. First observe that if $ps: \Theta \rightarrow [0, 1]$ has at least one value equal to 1, then $PL \circ MA(pl) = pl$. This follows directly from Theorem 5.2. Clearly in applying PL we lose information since mass functions have $2^n - 1$ degrees of freedom and pl-functions have only n. Our question is: to what extent do the plausibilities of a mass function on the singleton sets capture all the relevant information about the mass function? This is related to the question of how to define a decision rule for mass functions. One proposal is to use decisiveness (defined above, section 2.) to rank order the elements of a frame of discernment, the element with the largest decisiveness being the "correct" answer. We will show first that given a pl-function pl, the ordering given by pl and the ordering given by decisiveness of $MA(pl)$ are identical. Moreover, for any mass function, if it is consistent (the plausibility of some singleton is 1) then the rank order given by the plausibilities is identical with the rank order given by decisiveness.

Inherent to any pl-function is an ordering of Θ with respect to the values of the pl-function. Now observe that the mapping

$$MA\text{: pl-functions}(\Theta) \rightarrow \text{mass-functions}(\Theta)$$

yields an ordering of Θ given by decisiveness (see section 2.)

Theorem 6.1 *If $pl: \Theta \to [0,1]$ is a pl-function and $dec: \Theta \to [-1,1]$ is the decisiveness function of the mass-function $MA(pl)$ then the ordering of Θ given by pl and the ordering given by dec are identical.*

Proof: Assume $pl(a) \geq pl(b)$. Now from theorem 5.2,

$$dec(a) = pl(a)\left(\prod_{c \neq a}(1 - pl(c))\right) - 1 + pl(a)$$

and similarly for b. Rearranging terms then it follows that

$$dec(a) \geq dec(b)$$

if and only if

$$\prod_{c \neq a,b}(1 - pl(c))(1 - pl(b)) \geq \prod_{c \neq a,b}(1 - pl(c))(1 - pl(a))$$

and this latter inequality is clear from the assumption.

Now consider the mapping in the reverse direction

$$PL: \text{mass-functions}(\Theta) \to \text{pl-functions}(\Theta).$$

In general the ordering given by decisiveness and the ordering given plausibility are not equal. However in one important case they are equal. Define a mass function to be *consistent* if for some singleton the plausibility is equal to 1.

Theorem 6.2 *Suppose $m: 2^{\Theta} \to [0,1]$ is a consistent mass function. Then the ordering given by plausibility and the ordering given by decisiveness are identical.*

Proof: First observe that if $pls(a) = 1$ for some a then $spt(b) = 0$ for every b except possibily a. Thus it suffices to show that

$$dec(a) \geq dec(b)$$

for all $b \neq a$. But

$$dec(a) \geq dec(b)$$

if and only if

$$spt(a) + pls(a) \geq spt(b) + pls(b)$$

and the latter inequality is clear since $pl(a) = 1$.

Thus, with respect to consistent mass functions, and decisiveness as a decision rule, inference can be performed equally well with the computationally simpler measure given by plausibility.

7. Conclusions

In this paper we have considered mass functions generated from pl-functions defined from the statistics of features and objects. It is obvious that the mass assignments generated from pl-functions bear a great resemblance to probabilities on sets of independent events. For the examples given, this form is intuitively appealing as well as compact and easily analyzed. However, not all relationships between image features and their interpretations can be captured by the use of pl-functions in the way that we have defined the relationship between pl-functions and mass functions. We will conclude by considering two situations where this occurs.

First, a coarsening or refinement of the frame of discernment may be required. In this case the mass function on the refinement can not necessarily be generated by a pl-function over the refinement. For example, in the context of aerial photographs, a measure of rectangularity may discern between rectangular and non-rectangular objects but it is not appropriate to use this measure to distinguish between potentially rectangular objects (such as buildings or parking-lots). Therefore a single frame of discernment and related knowledge sources can not be used throughout the reasoning process and the system must be able manage mass functions of broader types than mentioned here.

Shafer [1982] suggests that in situations where the combination of mass functions produces a great deal of conflict the individual mass functions can be discounted then recombined. An example of this is uniform discounting which reduces the mass given to each proper subset and increases the mass given to Θ. If the discounting factor and conflict is large enough then the combined mass to each proper subset tends toward an "average" of individual mass functions. The given discounted mass function is not necessarily separable into the simple mass functions of the form described above, and thus the analysis using plausiblity functions does not apply.

Thus mass functions generated by pl-functions form only a proper subset of the mass functions which are appliciable in a general image understanding system. How-

ever they have much of the representational power normally associated with mass-functions and their simplicity and computational advantages make them very attactive in contexts where evidential reasoning and management of uncertainty is required.

Acknowledgements

We would like to thank Len Wesley, Al Hanson and Joey Griffith for many invaluble conversations concerning this paper. The work reported here was supported by the Air Force under AFOSR contract no. F49620-83-c-0099

References

P. Cheeseman (1983), "A method of computing generalized Bayesian probability values for expert systems", *Proc. Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, August 1983, pp. 198-202.

A. P. Dempster (1967), "Upper and lower probabilities induced by a multivalued mapping", *Annals of Mathematical Statistics*, vol. 38, 1967, pp. 325-339.

A. P. Dempster (1968), "A generalization of Bayesian inference", *Journal of the Royal Statistical Society*, Series B, vol. 30, 1968, pp. 205-247.

R.O. Duda and P.E. Hart (1973), *Pattern Classification and Scene Analysis*, New York:Wiley, 1973.

A. P. Dempster and A. Kong (1984), "Belief functions and communications networks", Report, Department of Statistics, Harvard University, Cambridge, MA, Nov. 1984.

O. D. Faugeras (1982), "Relaxation labeling and evidence gathering", *Proc. Conf. Pattern Recognition and Image Processing*, Las Vegas, June 1982, pp. 672-677.

R. A. Fisher (1922), "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society of London*, Series A, vol. 222, 1922, pp 309-368.

S. Geman and D. Geman (1984), "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images", *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. PAMI-6, no. 6, Nov. 1984, pp. 721-741.

I. J. Good (1962), "Subjective probability as a measure of a non-measurable set", in *Logic, Methodology and Philosophy of Science*, Nagel, Suppes and Tarski (eds), 1962.

A. Hanson, E. Riseman, J. Griffith, T. E. Weymouth (1984), "A methodology for the development of general knowledge-based vision systems",*Proc. IEEE Workshop on Principles of Knowledge Based Systems*, Denver, Colorado, Dec. 1984, pp. 159-170.

R. A. Hummel and S. W. Zucker (1983), "On the foundations of relaxation labeling processes", *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. PAMI-5, no. 3, May 1983, pp. 267-286.

R. A. Hummel and M. Landy (1985), "A Statistical Viewpoint on the Theory of Evidence", N.Y.U. Tech Report no. 194.

L. J. Kitchen (1980), "Relaxation applied to matching quantitative relational structures", *IEEE Trans. Syst. Man Cybern.*, vol. SMC-10, 1980, pp. 96-101.

L. J. Kitchen and A. Rosenfeld (1984), "Scene analysis using region-based constraint filtering", *Pattern Recognition*, vol. 17, no. 2, 1984, pp. 189-203.

H. E. Kyburg (1984), "Bayesian and non-Bayesian evidential updating", Tech. Report 139, Department of Computer Science, University of Rochester, Rochester NY, July 1984.

J. D. Lowrance (1982), "Dependency-graph models of evidential support", Ph.D. Thesis, University of Massachusetts, Amherst, 1982.

J. D. Lowrance, T. D. Garvey (1983), "Evidential reasoning: an implementation for multisensor integration", Tech Note 307, SRI Artificial Intelligence Center, December 1983.

S. Y. Lu, H. E. Stephanou (1984), "A set-theoretic framework for the processing of uncertain knowledge", AAAI-84, pp. 216-221.

G. Reynolds, D. Strahman, N. Lehrer (1985), *Converting Feature Values to Evidence*, DARPA Image Understanding Workshop, 1985, pp 331-339.

G. Reynolds, D. Strahman, N. Lehrer, L. Kitchen (1986), "Plausible Reasoning and the Theory of Evidence", UMASS COINS Tech. Rpt. April 1986.

A. Rosenfeld, R. A. Hummel and S. W. Zucker (1976), "Scene labeling by relaxation operations", *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, 1976, pp. 420-433.

G. Shafer (1973a), "A theory of statistical evidence", pp. 365-434 in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W. L. Harper and C. A. Hooker eds, vol. II.

G. Shafer (1976b), *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

G. Shafer (1981), "Constructive Probability", *Synthese*, vol. 48, 1981, pp. 1-60.

G. Shafer (1982), "Belief functions and Parametric Models", J. R. Statis. Soc. B (1982) 44, No3. pp. 322-352.

G. Shafer (1984), "Probability judgement in artificial intellegence and expert systems", Working Paper 165, School of Business, The University of Kansas, Lawrence, Kansas, Dec. 1984.

G. Shafer (1985), "Belief functions and possibility measures", *The Analysis of Fuzzy Information*, vol. 1, J. C. Bezdek, ed., CRC Press.

G. Shafer and A. Tversky (1983), "Weighing evidence: the design and comparison of probability thought experiments", *75th Anniversary Colloquium Series*, Harvard Business School.

L. Shastri and J. Feldman (1984), "Evidential Reasoning in Semantic Networks: a Formal theory", University of Rochester Tech. Report.

E.H. Shortliffe, "Computer-based medical consultations: MYCIN", New York, American Elsevier.

P. Smets (1978), "Un modele Mathematical-statistique simulant le processus du diagnostic medical", Doctoral Dissertation at the Free University of Brussels, Presses Universitaires de Bruxelles; also cited in G. Shafer (1982) .

C. A. B. Smith (1961), "Consistency in statisical inference and decision", *Journal of Royal Statisical Society*, Series B, vol 23, 1961, pp. 1-37.

C. A. B. Smith (1965), "Personal probability and statistical analysis", *Journal of the Royal Statisical Society*, Series B, vol. 128, 1965, pp. 469-499.

T. Strat (1984), "Continuous belief functions for evidential reasoning", AAAI-84, pp. 308-313.

L. Wesley (1983), "Reasoning about control: the investigation of an evidential approach", *Proc. Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, August 1983, pp. 203-210.

L. Wesley (1984), "Reasoning about control: an evidential approach", Tech. Report 324, SRI Artificial Intelligence Center, 1984.

L. Wesley (1985), Ph.D. Thesis, University of Massachusetts, Amherst, *in preparation.*

L. Wesley, A. Hanson (1985), "The application of an evidential-based technology to a high-level knowledge-based image interpretation system", to appear.

L. A. Zadeh (1978), "Fuzzy sets as a basis for a theory of possibility", Fuzzy Sets and Systems, 1, pp. 3-28.

L. A. Zadeh (1983), "The role of fuzzy logic in the management of uncertainty in expert systems", *Fuzzy Sets and Systems*, vol. 11, 1983, pp. 199-227.