Sensor Fusion Display Evaluation Using Information Integration Models in Enhanced/Synthetic Vision Applications

David C. Foyle
NASA Ames Research Center

ABSTRACT

Based on existing integration models in the psychological literature, an evaluation framework is developed to assess sensor fusion displays as might be implemented in an enhanced/synthetic vision system. The proposed evaluation framework for evaluating the operator's ability to use such systems is a normative approach: The pilot's performance with the sensor fusion image is compared to models' predictions based on the pilot's performance when viewing the original component sensor images prior to fusion. This allows for the determination as to when a sensor fusion system leads to: 1) poorer performance than one of the original sensor displays (clearly an undesirable system in which the fused sensor system causes some distortion or interference); 2) better performance than with either single sensor system alone, but at a sub-optimal (compared to model predictions) level; 3) optimal performance (compared to model predictions); or, 4) super-optimal performance, which may occur if the operator were able to use some highly diagnostic "emergent features" in the sensor fusion display, which were unavailable in the original sensor displays.

INTRODUCTION

Many different types of imaging sensors exist, each sensitive to a different region of the electromagnetic spectrum. Passive sensors, which collect energy emitted or reflected from a source, include television (visible light), night-vision devices (intensified visible and near-infrared light), passive millimeter wave sensors, and thermal imaging (infrared) sensors. Active sensors, in which objects are irradiated and the energy reflected from those objects is collected, include the various bands of radar (radio waves), such as x-band and millimeter wave.

These imaging sensors were developed because of their ability to increase the probability of identification or detection of objects under difficult environmental conditions. Because each sensor is sensitive to different portions of the spectrum, the resultant images contain different information when used under the same conditions. In order to present this information to an operator, image processing algorithms are being developed in many laboratories to "fuse" the information into a single coherent image containing information from more than one sensor (Toet, 1990; Pavel, Larimer & Ahumada, 1992). These displays are referred to as sensor fusion displays.

Sensor fusion displays are being considered in enhanced or synthetic vision systems for civil transport use. These displays would allow pilots to detect runway features and incursions during landing, and would aid in detecting obstacles and traffic in taxi (Foyle, Ahumada, Larimer & Sweet, 1992). Such sensor systems would allow continued operation in low-visibility weather conditions (i.e., the sensors would "see" through the fog).

Much of the role of enhanced and synthetic vision systems with sensor fusion can be characterized as a detection task for the pilot. These systems must allow the pilot to detect runway incursions by ground vehicles and by other aircraft, and to detect obstacles in taxi to the gate. Additionally, in order to complete an approach at an airport, the pilot must verify (detect) any of ten different visual references (see Table 1).

VISUAL REFERENCES TO COMPLETE APPROACH (FROM ACJ-OPS 1-3.20001 AND SIMILAR TO FAR 91.175)
THE APPROACH LIGHT SYSTEM
THE THRESHOLD
THE THRESHOLD MARKING
THE THRESHOLD LIGHTS
THRESHOLD IDENTIFICATION LIGHT
THE VISUAL GLIDE SLOPE INDICATOR
THE TOUCHDOWN ZONE OR TOUCHDOWN ZONE MARKINGS
THE TOUCHDOWN ZONE LIGHTS
THE RUNWAY LIGHTS

Table 1. Visual references required to be seen by the pilot at decision height to complete an approach under current FAA rules.

The work described in this paper was conducted to guide the development of such sensor fusion displays. An engineer developing such a system constantly reviews the resulting display and underlying algorithms on a subjective basis. More formal testing is also necessary. Suppose, for example, that two sensor sources individually allow the pilot to achieve 0.70 probability of runway incursion detection under some particular environmental conditions. What, then, is the expected probability of runway incursion detection when the two sensors are combined according to some image processing technique? If observed runway incursion detection improves with a sensor fusion system to 0.80, is that a large improvement, or should one actually expect more? The ability to answer these types of questions can lead to a better human-machine system in two ways: Proposed sensor integration hardware and software can be evaluated both relatively, by determining which sensors and algorithm combinations are better than others, and absolutely, by comparing system (pilot/display) performance to theoretical expectations.

INFORMATION INTEGRATION MODELS

Previous work has been conducted on the topic of how operators integrate the information from multicomponent auditory signals, from the visual and auditory senses, and from multiple observations over time (Green, 1958; Craig, Colquhoun & Corcoran, 1976; Green & Swets, 1966/1974). These models all predict operator integration performance as a function of the operator's performance with the individual stimuli

comprising the integration task. Two classes of models have been developed: Decision combination models and observation integration models (Swets, 1984). The decision combination models assume that in the integration task the operator makes an individual decision about each aspect of the combined display and then combines those decisions to yield one final decision. At the time of the final decision, only the previous decisions are available, and not the information that led to the individual decisions. The observation integration models, on the contrary, assume that the operator does have access to that information. The internal representations of the individual observations (e.g., likelihood ratios) are then combined, yielding only one decision.

The simplest version of a decision combination model is the probability summation, or statistical summation, model. It is derived from the independence theorem of probability theory and was first proposed by Pirenne as a perceptual model (Pirenne, 1943; Swets, 1984). In its simplest form, the two information sources are assumed to be independent and uncorrelated. It states that performance with a complex stimulus is predictable from the performance with the individual stimuli according to the following equation:

$$p_{12} = p_1 + p_2 - p_1 p_2 \tag{1}$$

where p_1 and p_2 represent detection probabilities for the two stimuli presented in isolation, and p_{12} is the detection probability when both stimuli are available.

The most cited version of the observation integration model is derived from the theory of signal detectability and was originally proposed by Green (1958). As in Pirenne's (1943) model, in its most simple form, the information from the two sources is also assumed to be independent and uncorrelated. The model is stated in terms of the sensitivity measure, d':

$$d'_{12} = \sqrt{(d'_1)^2 + (d'_2)^2} \tag{2}$$

where d'_1 and d'_2, and d'_{12}, respectively, represent performance with the two stimuli presented in isolation, and when both stimuli are available.

Swets has noted that the statistical summation model fits simple detection data fairly well when the observed detection probabilities are corrected for chance success (Swets, 1984). Similarly, in the experiments in which it has been applied, the observation integration model well represents the data.

The two integration models presented here have been incorporated into the development of a framework which can be used to evaluate combined human-machine performance for sensor fusion displays.

PROPOSED EVALUATION FRAMEWORK

A sensor fusion display typically refers to the combined image display resulting from the application of an image processing technique on two or more individual sensor

images. The proposed framework for evaluating the operator's ability to use such systems is a normative approach: The operator's performance with the sensor fusion display can be compared to performance on the individual sensor displays comprising that display and to various optimal models of integration.

Typically, as the environmental conditions change in which the individual sensor operates, so does the information content of that image. The information content of the image can be "scaled" by the operator's ability to perform a target identification or discrimination task (e.g., detecting a runway incursion). One would expect task performance with a sensor fusion display formed from two low information content (hence poor performance) images, to still be relatively poor. Similarly, two high information content (high performance) sensor images should yield good performance

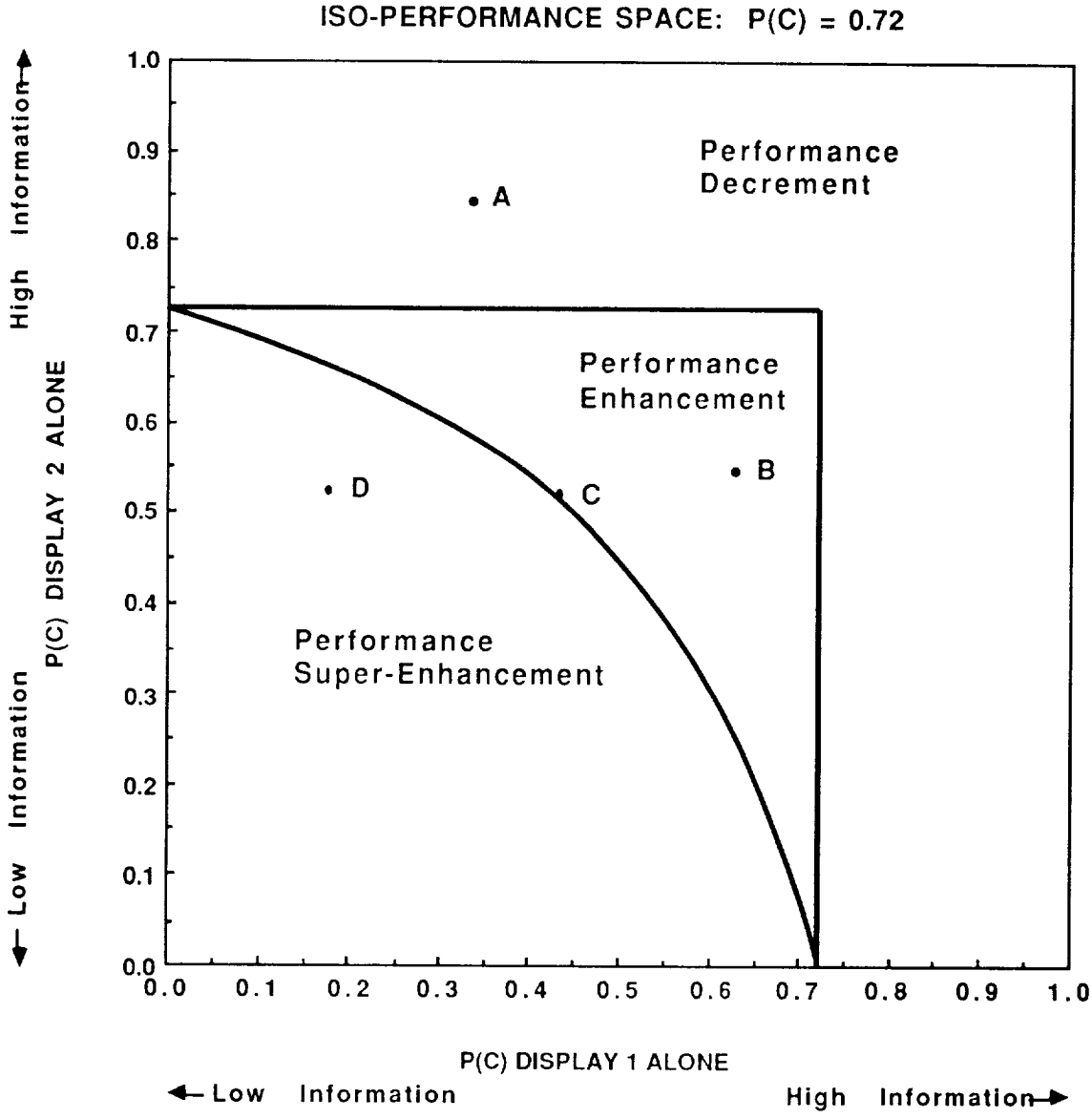

Fig. 1. A proposed evaluation framework for sensor fusion displays. All data points represent P(C) = 0.72 for the dual display or sensor fusion display task.

when combined into a sensor fusion display. Assuming that there was some independent information in the two individual sensor images, one would also expect performance with the sensor fusion display to be better than with either of the two individual sensors alone. This results in a 3-dimensional performance space: Performance with the sensor fusion image is a function of the performance levels associated with the two individual sensor images.

Figure 1 shows part of this performance space associated with a sensor fusion display. The abscissa and the ordinate result from the stimulus-performance scaling for Sensor Display 1 and Sensor Display 2, respectively, when viewed by an operator in isolation. The figure shows the iso-performance horizontal "slice" through the

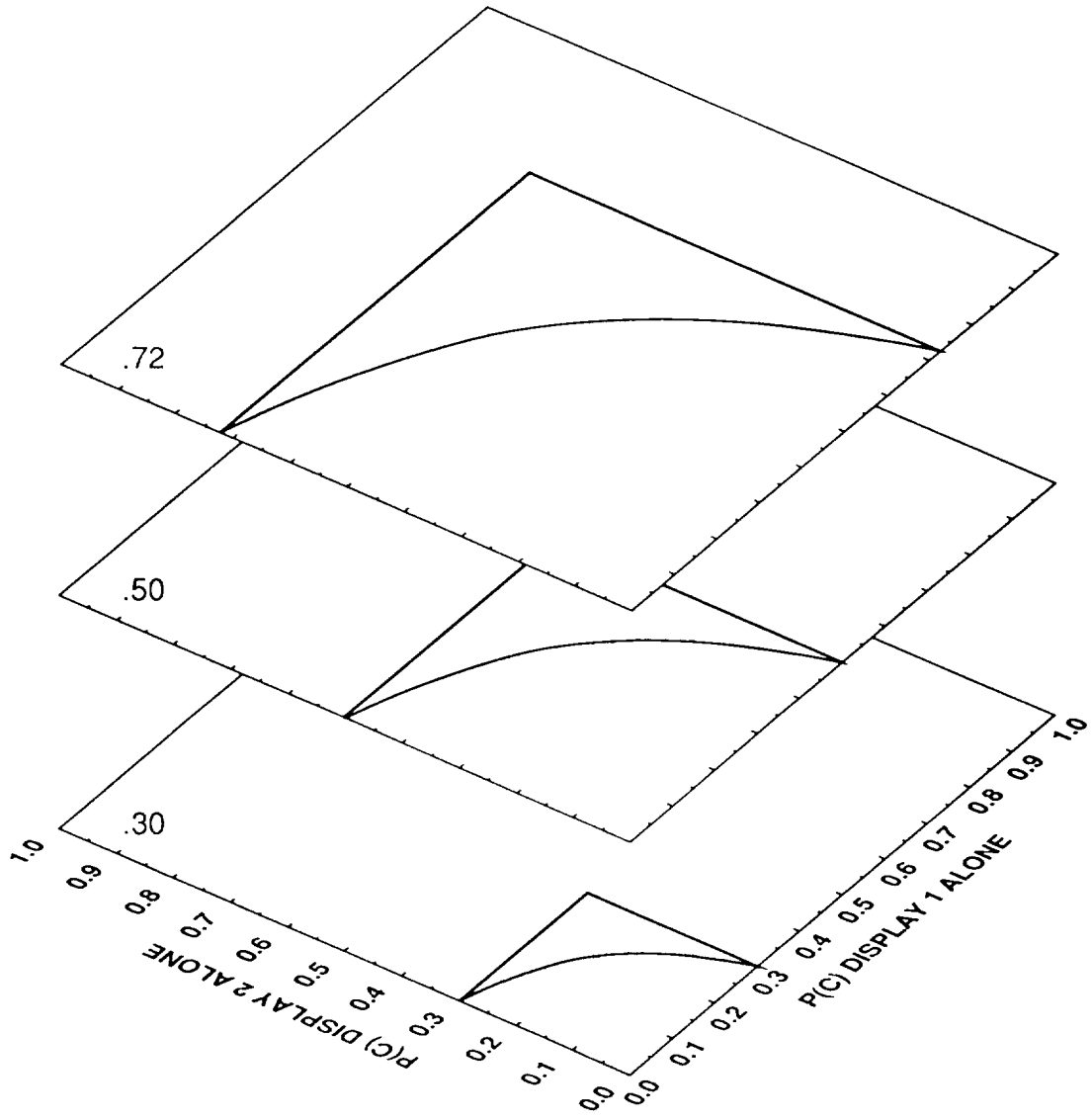

Fig. 2. Three example horizontal slices through the 3-dimensional performance space. The value on each overlay represents the performance level, in P(C), for the sensor fusion display task.

3-dimensional space in which all performance data points represent 0.72 (corrected for chance) detection probability using a sensor fusion display. As noted above, the actual performance space is 3-dimensional and is represented in Figure 2 by similar-appearing "slices" for three example performance levels.

Because the sensor fusion display data are plotted as iso-performance slices, data points near the origin represent better performance than away from the origin. For the same level of performance, a data point near the origin represents a condition in which very little information was available in the two displays, whereas a data point away from the origin refers to a condition in which relatively more information was available in the separate displays. Thus, for a given resultant sensor fusion performance level (i.e., "horizontal slice") data points near the origin represent better sensor fusion displays.

In these figures and all remaining references, P(C) refers to the proportion of correct responses with a correction for chance applied. A correction for chance is necessary when measuring performance in P(C) units because the integration models require that a performance level of zero be associated with the operator receiving no information from the display. No such correction is necessary when measuring performance in d' units since d' = 0 refers to chance performance.

As can be seen from the two figures, the sensor fusion performance space can be divided into three separate areas, Performance Decrement, Performance Enhancement, and Performance Super-Enhancement, each with unique interpretations if data points lie in those areas. The two right-angle lines dividing the Performance Decrement and Performance Enhancement areas are determined by the horizontal and vertical lines crossing the axes at the level of performance [P(C) = 0.72 in Figure 1] for the sensor fusion display. The smooth curves separating the Performance Enhancement and Performance Super-Enhancement areas are the predictions of the statistical summation model (see eq. 1) where p_{12} = 0.72 in Figure 1 and 0.30, 0.50, and 0.72 in Figure 2. Because these two models predict optimal performance (that is, they both assume ideal observers with no memory limitations, etc., with independent and uncorrelated information in the separate displays) their predictions can be used as an upper bound against which to measure integration performance. The interpretation of the data points falling into the three areas is best illustrated by example.

Performance decrement

Suppose under a given environmental condition, an operator achieved runway incursion detection performance of P(C) = 0.33 when viewing Sensor 1 in isolation and P(C) = 0.84 when viewing Sensor 2 in isolation. When these two sources are both available (separately on two monitors, or fused on a single monitor according to a sensor fusion algorithm) to the operator and performance is P(C) = 0.72, the resultant data point would be the one labeled "A" in Figure 1. Obviously, in this situation, the sensor fusion display has not improved the pilot's overall runway incursion detection performance. In fact, performance in the sensor fusion display case has now decreased to only P(C) = 0.72, whereas previously the operator was able to use Sensor 1 in isolation and reach P(C) = 0.84 performance. Such a performance decrement could be the result of the deletion of necessary information by the sensor fusion algorithm, or could represent a cognitive limitation on the part of the pilot.

Performance enhancement

Data point "B" in Figure 1 would result if P(C) = 0.72 performance obtained using the sensor fusion display, when Sensors 1 and 2 yielded P(C) = 0.63 and P(C) = 0.55 in isolation. In this case, performance has improved, since the pilot is now doing better with the sensor fusion display (0.72) than with either of the two sources alone (0.55, 0.63). However, the two models of information integration predict a larger improvement in this case. Thus, for data points falling in this region, there is performance improvement, but one would expect more. Data point "C", lying on the statistical summation model curve, represents optimal integration performance, in which sensor fusion display performance of 0.72 is expected if performance on Sensor 1 were 0.42 and Sensor 2 performance was 0.52.

Pilot detection performance occurring in this region would occur when some of the information in the two sources is redundant (correlated and not independent), or when the sensor fusion algorithm integrates the information suboptimally. The statistical summation model (as well as the observation integration model) can be viewed as an upper limit of integration: It assumes that the information in the two sources is independent and non-redundant, and does not assume any decrease in performance due to the limits of cognitive processes (i.e., memory, workload, or suboptimal strategies).

Performance super-enhancement

Data point "D" would result when the individual runway incursion detection performance for the two sensors alone was P(C) = 0.17 and P(C) = 0.52 and sensor fusion display performance was P(C) = 0.72. Data points falling in this region between the model prediction and the origin represent improved performance that is better than is predictable from the model. That is, when the sensor fusion display is viewed, some new, previously unusable, information emerges which results in much better performance.

The random-dot stereogram display can be thought of as an example of a sensor fusion display that has these properties (Julesz, 1971). In these displays, random dots are offset differentially yielding a perception of an object in the third dimension. In such a stereogram there is no information whatsoever in the individual halves of the stereogram, but only in differences between the two displays. The object is observable only by stereoscopically fusing the two halves of the stereogram or analytically determining the differences. In fact, if one conducted an experiment in which subjects had to state the "floating" shape, one would obtain chance performance when viewing only one stereogram half and perfect performance when both stereogram pairs are viewed. This represents Performance Super-Enhancement because based on chance performance with the stereogram halves, one would conclude that they contain no information. This would lead one to predict chance performance when both halves are available, which obviously is not the case. Conditions in which Performance Super-Enhancement occurs could be capitalized upon to produce useful sensor fusion techniques. The proposed evaluation framework provides for the ability to recognize and quantify such conditions.

Evaluation framework implementation

In order to evaluate human performance with a sensor fusion system using the proposed evaluation framework, the following steps must be taken:

Performance scaling of Sensor 1. Determine the psychometric function relating task performance (e.g., runway incursion detection, runway lights detection) to the environmental conditions of interest. For example, infrared imagery is degraded by increasing atmospheric moisture. The information content of each sensor image varies with the environmental conditions, and in a sense, this scaling estimates the amount of information available to the operator with Sensor 1 alone under those conditions.

Performance scaling of Sensor 2. Similar to Sensor 1.

Performance with sensor fusion display. For various combinations of environmental or sensor conditions previously evaluated in isolation, determine task performance using the proposed fusion algorithm and associated display.

Performance with operator integration. As in the sensor fusion evaluation phase, determine task performance with both sensors but with either two displays or a split screen. This step acts as a control condition, and essentially allows the operator to integrate the information from the two sensors. A sensor fusion algorithm should yield better task performance than when the operator uses two displays or a split-screen display.

SENSOR FUSION EVALUATION: FOYLE (1992)

To illustrate how the evaluation framework would be used the results from an experiment are briefly presented. In an experiment reported in Foyle (1992), subjects had to integrate the information in two sensor displays to detect a target. As an experimental convenience, combinations of separate sensor sources yielding an iso-performance level [P(C) = 0.72] of integration performance were determined (with both sensor sources available on multiple screens, analogous to performance with a sensor fusion display). These combinations were then plotted on the evaluation framework graph.

Figure 3 shows combinations of the individual sensor sources, in P(C) units as scaled by P(C) psychometric functions, yielding P(C) = 0.72 dual-display (sensor fusion) performance. The two curves represent predictions of the two optimal integration models (statistical summation and observation integration) as described by the equations shown in the figures. For illustration purposes, note the right-most (also lower-most) data point for subject 4. That data point shows that P(C) = 0.72 detection performance obtained when viewing two sensor displays simultaneously: A Sensor 1 image display which yielded P(C) = 0.60 probability of detection alone, and a Sensor 2 image display which yielded P(C) = 0.36 probability of detection alone.

Analyzing the results of this experiment using this method, Foyle (1992) concluded that ten of the eighteen data points in Figure 3 lie in the triangular "performance enhancement" region when plotted onto the evaluation framework graph. For those conditions, the subjects were able to integrate the images from the two displays and performed better than when only one of those displays was available. The conditions that led to integration occurred when Sensor Display 1 yielded moderate detection performance (approximately P(C) = 0.50 in Figure 3). When a low-quality image (yielding about P(C) = 0.30) was presented as Sensor Display 1, the images in Sensor Display 2 were required to be of very high-quality in order to yield P(C) = 0.72 with both displays. In fact,

they were of such high quality that when presented in isolation, they would have yielded performance of P(C) = 0.80 or 0.90. The subjects would have done better in those conditions if they had simply ignored the low-quality images on Sensor Display 1 and based their responses only on the images on Sensor Display 2. (Graphically, that would have forced the data points onto the horizontal straight line in Figure 3.)

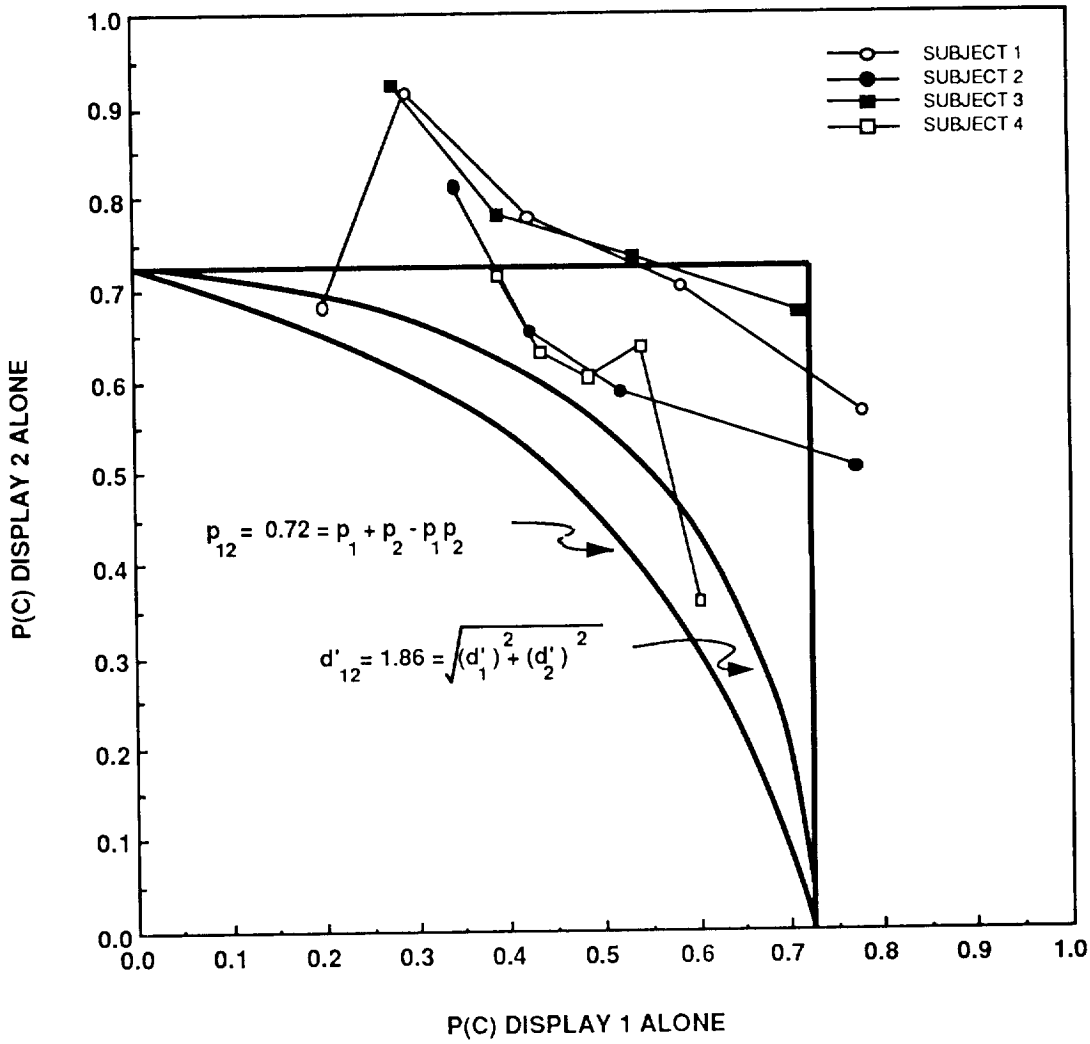

Fig. 3. Experimental data from Foyle (1992), in corrected-for-chance P(C), overlaid on the proposed evaluation framework. All data points in this "horizontal slice" through the 3-dimensional space represent P(C) = 0.72 detection probability performance.

These data were explained by a model in which subjects always give equal weight to the information in the two displays despite the image quality level. The effect may be similar to that noted by Tversky and Kahneman (1974) in which subjects weighted obviously irrelevant information equally with relevant information. The conditions under which subjects are able to integrate display information, and those that do not facilitate, and actually decrease performance clearly warrant more investigation. As stated earlier, the statistical summation and observation integration models can be viewed as an upper bound to normal (not Performance Super-Enhancement) information integration. In this

particular experiment, the model predictions were not only an upper bound on performance in general, but in fact were appropriate predictions since the information in the dual-display condition was independent and uncorrelated. The models' failure to predict the data establishes the existence of the subjects' cognitive limitations in this particular task.

CONCLUSIONS AND SUMMARY

For a sensor fusion display in an enhanced or synthetic vision system, much of what the pilot must do with the system is to detect traffic and detect certain visual references in order to complete an approach and land. The evaluation framework described in this paper allows system engineers and researchers to evaluate pilot-in-the-loop performance with the sensor fusion algorithms and display against a theoretical optimal benchmark. By using such a benchmark, the system engineer can ensure that the important features available in the sensor imagery prior to fusion are preserved.

In summary, the evaluation framework developed herein has been demonstrated to be a useful tool to evaluate pilot's ability to extract information from a sensor fusion display or to integrate information from two displays. The techniques discussed allow the evaluation of sensor fusion displays by comparing sensor fusion display performance to the predictions of existing optimal integration models and to multiple display presentations. This evaluation allows the human factors engineer to recognize in an absolute sense, as well as relative, whether the proposed sensor fusion display does what it was designed to do: integrate the sensor information and present it well.

REFERENCES

Craig, A., Colquhoun, W.P. and Corcoran, D.W.J. (1976). Combining evidence presented simultaneously to the eye and the ear: A comparison of some predictive models. *Perception & Psychophysics, 19*, 473-484.

Foyle, D.C. (1992). Proposed evaluation framework for assessing operator performance with multisensor displays. SPIE Volume 1666: *Human Vision, Visual Processing, and Digital Display III*, 514-525.

Foyle, D.C., Ahumada, A.J., Larimer, J. and Sweet, B.T. (1992). Enhanced/synthetic vision systems: Human factors research and implications for future systems (Paper 921968). *Enhanced Situation Awareness Technology for Retrofit and Advanced Cockpit Design*, SAE SP-933: Conference Proceedings of the SAE Aerotech '92 Meeting (Human Behavioral Technology/Aerospace Technologies Activity).

Green, D.M. (1958). Detection of multiple component signals in noise. *Journal of the Acoustical Society of America, 30*, 904-911.

Green, D.M. and Swets, J.A. (1974). *Signal Detection Theory and Psychophysics*, Krieger, New York, 1974. (Originally published 1966 by Wiley, New York.)

Julesz, B. (1971). *Foundations of Cyclopean Perception*, University of Chicago Press, Chicago.

Pavel, M., Larimer, J., and Ahumada, A. (1992). Sensor fusion for synthetic vision. *SID International Symposium Digest of Technical Papers, 23,* 475-478.

Pirenne, M.H. (1943). Binocular and uniocular thresholds in vision. *Nature, 152,* 698-699.

Swets, J.A. (1984). Mathematical models of attention. In *Varieties of Attention,* R. Parasuraman and D.R. Davies (Eds.), 183 - 242, Academic Press, New York.

Toet, A. (1990). Hierarchical image fusion. *Machine Vision and Applications, 3,* 1-11.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty. *Science, 185,* 1124-1130.