Physics-Based Integration of Multiple Sensing Modalities for Scene Interpretation

N. NANDHAKUMAR, SENIOR MEMBER, IEEE, AND J. K. AGGARWAL, FELLOW, IEEE

Invited Paper

The fusion of multiple imaging modalities offers many advantages over the analysis, separately, of the individual sensory modalities. In this paper we present a unique approach to the integrated analysis of disparate sources of imagery for object recognition. The approach is based on physics-based modeling of the image generation mechanisms. Such models make possible features that are physically meaningful and have an improved capacity to differentiate between multiple classes of objects. We illustrate the use of physics-based approach to develop multisensory vision systems for different object recognition application domains. The paper discusses the integration of different suites of sensors, the integration of image-derived information with model-derived information, and the physics-based simulation of multisensory imagery.

I. INTRODUCTION

It is well known that the human visual system extracts a great deal of information from a single gray level image. This fact motivated computer vision researchers to devote much of their attention to analyzing isolated gray scale images. However, research in computer vision has made it increasingly evident that formulation of the interpretation of a single image (of a general scene) as a computational problem results in an underconstrained task. Several approaches have been investigated to alleviate the ill-posed nature of image interpretation tasks. The extraction of additional information from the image or from other sources, including other images, has been seen as a way of constraining the interpretation [1]. Such approaches may be broadly grouped into the following categories: 1) the extraction and fusion of multiple cues from the same image, e.g., the fusion of multiple shape-from-X methods (e.g., shape from shading, shape from texture), 2) the use of multiple views of the scene, e.g., stereo, and more recently 3) the fusion of information from different modalities of sensing, e.g., infrared and laser ranging.

Various researchers have referred to each of the above approaches as multisensory approaches to computer vision. The order in which the above approaches have been listed indicates, approximately, the chronological order in which these methods have been investigated. The order is also indicative of the increasing amount of additional information that can be extracted from the scene and which can be brought to bear on the interpretation task.

Past research in computer vision has yielded analytically well defined algorithms for extracting simple information (e.g., edges, two-dimensional (2-D) shape, stereo range, etc.) from images acquired by any one modality of sensing. When multiple sensors, multiple processing modules, and/or different modalities of imaging are to be combined in a vision system, it is important to address the development of 1) models relating the images of each sensor to scene variables, 2) models relating sensors to each other, and 3) algorithms for extracting and combining the different information in the images.

The choice of a computational framework for a multisensory vision system depends on the application task. Several computational paradigms have been employed in different recent multisensory vision systems. The paradigms can be categorized as: 1) statistical, 2) variational, 3) artificial intelligence (AI), and 4) phenomenological (physics-based) approaches. Statistical approaches typically involve Bayesian schemes which model multisensory information using multivariate probability models or as a collection of individual (but mutually constrained) classifiers/estimators. These schemes are appropriate when the domain of application renders probabilistic models to be intuitively natural forms of models of sensor performance and the state of the sensed environment. An alternative, deterministic, approach is based on variational principles wherein a criterion functional is optimized. The criterion functional

Manuscript received May 29, 1996; revised September 6, 1996. N. Nandhakumar's work was supported by a RADIUS seed contract from Lockheed-Martin, AFOSR Contract F49620-93-C-0063, and ARPA Contract F33615-94-C-1529. J. K. Aggarwal's work was supported by Army Research Office Contracts DAAH-94-G-0417 and DAAH 049510494.

N. Nandhakumar is with Electroglas, Inc., Santa Clara, CA 95054 USA (e-mail: nnandhak@electroglas.com).

J. K. Aggarwal is with the Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712-1084 USA (e-mail: jka@uts.cc.utexas.edu).

Publisher Item Identifier S 0018-9219(97)00641-5.

implicitly models world knowledge and also explicitly includes constraints from multiple sensors. Adoption of this approach results in an iterative numerical relaxation approach which optimizes the criterion functional.

The complexity of the task sometimes precludes simple analytical formulations for scene interpretation tasks. Models relating the images of each sensor to scene variables, models relating sensors to each other, and algorithms for extracting and combining the different information in the images usually embody many variables which are not known prior to their interpretation. This necessitates the use of heuristic and empirical methods for analyzing the images. The development of complex interpretation strategies and knowledge representational mechanisms for using such methods has been intensively researched in the field of AI. Many of these ideas can be employed in the design of a multisensory vision system.

Most of the research (to date) in multisensory computer has adopted either a statistical-, variational-, or AI-based approach. Only very recently has research been directed at using phenomenological or physics-based models for multisensory vision. These models are based on physical laws, e.g., the conservation of energy. Such models relate each of sensed signals to the various physical parameters of the imaged object. The objective is to solve for the unknown physical parameters using the known constraints and signal values. The physical parameters then serve as meaningful features for object classification.

Denote sensed information as s_i. Each imaging modality (viz. physical sensor) may yield many types of sensed information s_i. For example, we may have $s_1 =$ "thermal intensity," $s_2 =$ "stereo range," $s_3 =$ "visual intensity," $s_4 =$ "visual edge strength," etc. Let $I_{s_i}(x, y)$ denote the value of sensed information s_i at any specified pixel location (x, y). For the sake of brevity, I_{s_i} will be used instead of $I_{s_i}(x, y)$ in the following. Each source of information is related to object parameters, b_j, and ambient scene parameters, c_j, via a physical model of the following general form

$$I_{s_i} = g_i(b_1, \cdots, b_N, c_1, \cdots, c_M) \tag{1}$$

where N and M are the number of object and scene parameters, respectively. Note that for each g_i, only a subset of the entire set of parameters has nonzero coefficients. Examples of b_j include visual reflectance of the surface, relative surface orientation, material density, and surface roughness. Examples of c_j include ambient temperature, direction of illumination, intensity of solar insolation, and ambient wind speed. The functional form of g_i depends on the sensor used and the specific radiometric, photometric, or projective principle that underlies image formation in that sensor. In addition to the above, various natural laws describe the physical behavior of the objects, e.g., principles of rigidity and the law of the conservation of energy. These lead to additional (known) constraints of the following general form

$$L_k(b_1, \cdots, b_N, c_1, \cdots, c_M) = 0. \tag{2}$$

The general problem is to use the functional forms (1) and (2) established above to derive a feature f that depends only the object properties, b_j, and are independent of scene variables, c_j. In some cases, it may be possible to solve for a specific object property c_j itself. If this property is a distinguishing and scene-invariant property of the object, such as material density or thermal capacitance, then this property can be used as the feature, f, for recognition. Note that in general, the equations are nonlinear, and hence solving them is not straightforward. Also, it may be possible to specify a larger number of equations than required, thus leading to an overconstrained system. An error minimization approach may then be used to solve for the unknowns. Alternatively, an overconstrained system supports the derivation of algebraic invariants [2] that are functions of image information and physical properties of objects, and invariant to scene conditions.

The physics based approach has been sucessfully adopted for a number of scene interpretation tasks. Section II provides an overview of several of these applications—and thus illustrates the usefulness of this paradigm as a general approach for the interpretation of multisensory imagery. Section III describes the integration of multisensory information for computing features based on physics-based models where some of the information is obtained form the image, and some of it is obtained from the model of the hypothesized object. Another important issue related to the interpretation of multisensory imagery is the simulation of multisensory imagery. A unified scheme for generating imagery as would be sensed by different co-boresighted sensors is described in Section IV. Section V contains concluding remarks.

II. Integration of Different Sensing Modalities

The physics-based approach has been used to address, successfully, a number of object recognition tasks that rely on the analysis of multisensory imagery. We provide below an overview of several of these approaches to illustrate this paradigm. We describe the integration of infrared and visual imagery for classifying objects occurring in outdoor scenes into broadly defined classes, such as vehicles, buildings, pavement, and vegetation. Next, we describe the integration of underwater acoustic and visual imagery for the classification of the imaged sea floor into two types—sediment or manganese deposits. We then describe the interpretation of different channels of polarized radiation sensed by optical cameras, and by a synthetic aperture radar (SAR) system. We end the section with brief discussions on the fusion of some other sensing modalities such as different channels of color information, and optical and radar imagery.

A. Integration of Infrared and Optical Imagery

The interpretation of thermal and visual imagery of outdoor scenes using a physics-based approach has been described in [3]–[6]. The approach is based on a model of energy exchange between the imaged surface and the

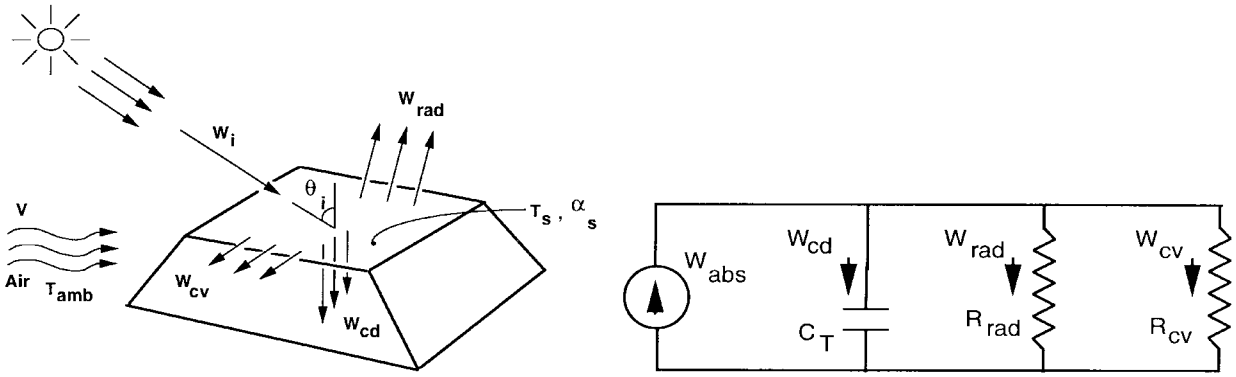


Fig. 1. (a) Energy exchange at the surface of an object in an outdoor scene and (b) equivalent thermal circuit for the surface of the object.

environment (including the cameras). This approach allows the estimation of internal object properties of the object such as its ability to sink/source radiation, i.e., its thermal capacitance. These estimates of internal object properties are physically meaningful and powerful features for classifying objects into broadly defined classes such as buildings, vegetation, vehicles, and roads. These features are tolerant to changes in viewing direction, illumination, surface coating, and ambient conditions. The approach can therefore be used in different types of outdoor autonomous navigation systems. It is interesting to note that the perceptual systems in certain snakes combine thermal and visual imaging senses to produce such a map of heat sinks and sources [7]–[9].

An overview of this approach is presented. The analysis of two types of imagery is discussed: 1) single set of multisensory data and 2) a temporal sequence of multisensory data. The model used for integrating thermal and visual imagery of outdoor scenes is based on the principle of conservation of energy. At the surface of the imaged object [Fig. 1(a)] energy absorbed by the surface equals the energy lost to the environment.

$$W_{\rm abs} = W_{\rm lost}. \tag{3}$$

Energy absorbed by the surface is is given by

$$W_{\rm abs} = W_I \cos\theta_I \alpha_s \tag{4}$$

where W_I is the incident solar irradiation on a horizontal surface and is given by available empirical models (based on time, date and latitude of the scene) or by measurement with a pyrheliometer, θ_I is the angle between the direction of irradiation and the surface normal, and α_s is the surface absorptivity which is related to the visual reflectance ρ_s by $\alpha_s = 1 - \rho_s$. Note that it is reasonable to use the visual reflectance to estimate the energy absorbed by the surface since approximately 90% of the energy in solar irradiation lies in the visible wavelengths [10]. A simplified shape-from-shading approach is used to compute $\cos\theta_I$ and α_s from the visual image and is described in detail in [3]. The energy lost by the surface to the environment is given by

$$W_{\rm lost} = W_{\rm cd} + W_{\rm cv} + W_{\rm rad} \tag{5}$$

where $W_{\rm cv}$ denotes the heat convected from the surface to the air which has temperature $T_{\rm amb}$ and velocity V, $W_{\rm rad}$ is the heat lost by the surface to the environment via radiation and $W_{\rm cd}$ denotes the heat conducted from the surface into the interior of the object. The radiation heat loss is computed from

$$W_{\rm rad} = \epsilon\sigma\left(T_s^4 - T_{\rm amb}^4\right) \tag{6}$$

where σ denotes the Stefan–Boltzman constant, T_s is the surface temperature of the imaged object, and $T_{\rm amb}$ is the ambient temperature.

The surface temperature is computed from the thermal image based on an appropriate model of radiation energy exchange between the surface and the infrared camera [3]. The resulting relationship between the surface temperature and image gray level produced by the infrared camera is of the form

$$\epsilon \int_{\lambda_1}^{\lambda_2} \frac{C_1}{\lambda^5(\exp(C_2/\lambda T_s) - 1)} d\lambda = K_a L_t + K_b \tag{7}$$

where C_1, C_2, K_a, and K_b are known constants, T_s is the surface temperature of the imaged object, λ is the wavelength of energy, $\lambda_1 = 8$ μm and $\lambda_2 = 12$ μm for the specific camera being used, ϵ is the thermal emissivity of the surface, and L_t is the gray level of a pixel in the thermal image.

The convected heat transfer is given by

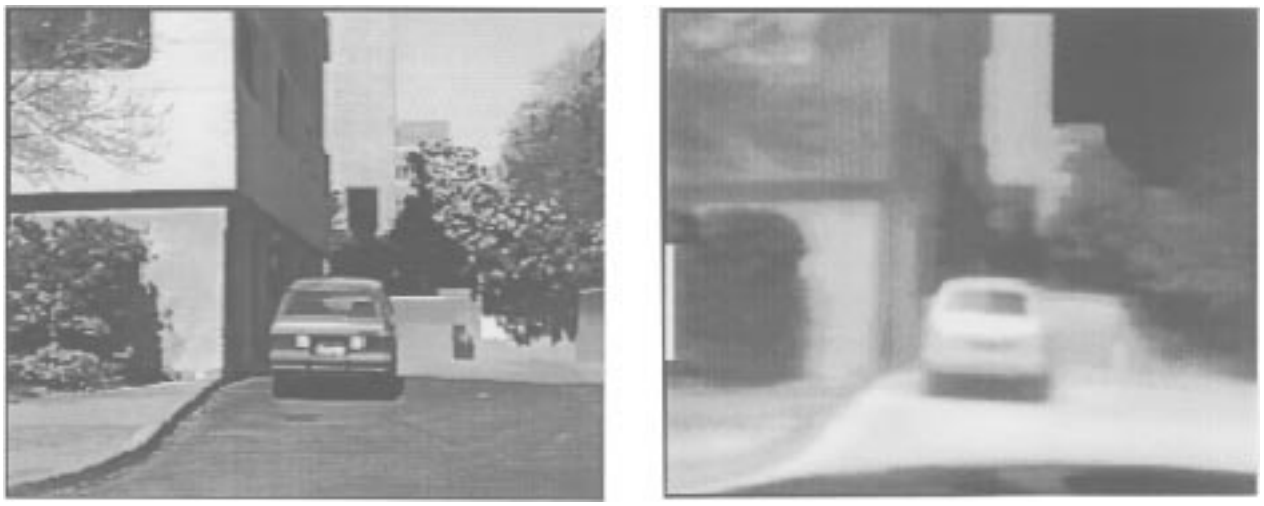
$$W_{\rm cv} = h(T_s - T_{\rm amb}) \tag{8}$$

where h is the average convected heat transfer coefficient for the imaged surface. For mixed air flow conditions this coefficient may be computed by

$$h = A_1 s^{\frac{4}{5}} L^{-\frac{1}{5}} - A_2 L^{-1} \tag{9}$$

where A_1 and A_2 are known thermophysical constants of the surrounding air [10], s is the wind speed, and L is the characteristic length of the imaged object. In the existing method, the wind speed is measured by an anemometer and a value of 1 m is assumed for L

Considering a unit area on the surface of the imaged object, the equivalent thermal circuit for the surface is



(a) (b)

Fig. 2. (a) Visual image of a scene and (b) thermal image of the scene.

Table 1 Thermal Capacitance of Objects Commonly Imaged in Outdoor Scenes. A Useful and Physically Menaingful Feature for Object Recognition.

Object	*Thermal Capacitance* ($\times 10^{-6}$ J/K)
Asphalt Pavement	1.95
Concrete Wall	2.03
Brick Wall	1.51
Wood (Oak) Wall	1.91
Granite	2.25
Automobile	0.18

shown in Fig. 1(b). C_T is the lumped thermal capacitance of the object and is given by

$$C_T = DVc$$

where D is the density of the object, V is the volume, and c is the specific heat. The resistances are given by

$$R_{\rm cv} = \frac{1}{h} \quad \text{and} \quad R_{\rm rad} = \frac{1}{\epsilon\sigma\left(T_s^2 + T_{\rm amb}^2\right)\left(T_s + T_{\rm amb}\right)}.$$

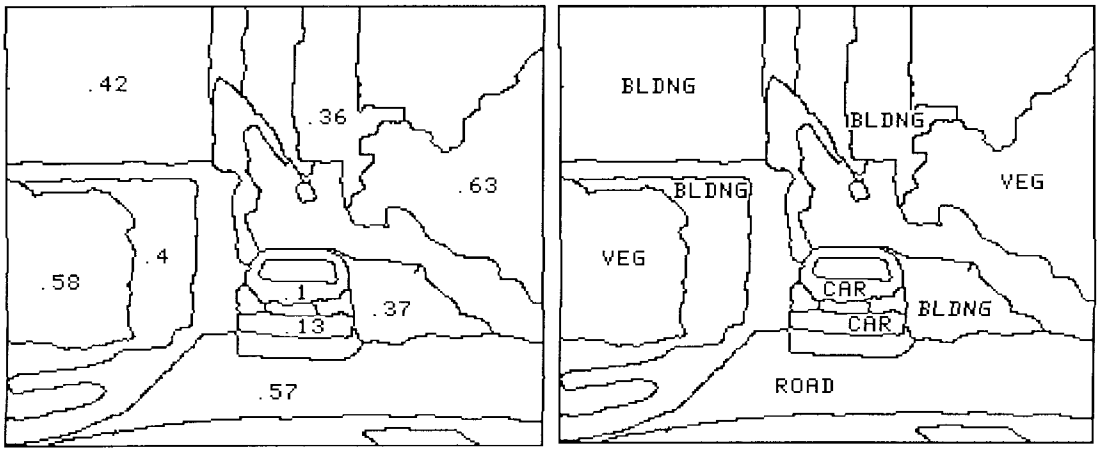
1) Analyzing a Single Set of Multisensor Data: It is clear from Fig. 1(b) that the conduction heat flux $W_{\rm cd}$ depends on the lumped thermal capacitance C_T of the object. A relatively high value for C_T implies that the object is able to sink or source relatively large amounts of heat. An estimate of $W_{\rm cd}$, therefore provides us with a relative estimate of the thermal capacitance of the object, albeit a very approximate one. Table 1 lists values of C_T of typical objects imaged in outdoor scenes. The values have been normalized for unit volume of the object.

Note that the thermal capacitance for walls and pavements is significantly greater than that for automobiles and hence $W_{\rm cd}$ may be expected to be higher for the former regions. Plants absorb a significant percentage of the incident solar radiation which is used for photosynthesis and also for transpiration. Very little of the absorbed radiation is convected into the air. Therefore, the estimate of the $W_{\rm cd}$ will be almost as large (typically 95%) as that of the absorbed heat flux. Thus $W_{\rm cd}$ is useful in estimating the object's ability to sink/source heat radiation, a feature shown to be useful in discriminating between different classes of objects. However, in order to minimize the feature's dependence on differences in absorbed heat flux, a normalized feature was defined to be the ratio $R = W_{\rm cd}/W_{\rm abs}$.

The steps involved in evaluating this internal object property from the thermal and visual image pair of a scene are as follows. The registered thermal and visual image pair is segmented. The simplified shape-from-shading approach discussed in [3] yields the surface reflectance for each region, and it yields the value of $\cos\theta_I$ at each pixel. The thermal image provides an estimate of surface temperature T_s at each pixel by using a table of values of L_t generated by (7) for different values of T_s and a fast table-lookup scheme. A value of 0.9 is assumed for the surface emissivity of all objects. Equations (3)–(6), (8), and (9) are applied at each pixel to estimate $W_{\rm cd}$ and thence the ratio $R = W_{\rm cd}/W_{\rm abs}$.

The above approach was tested on real data. Fig. 2(a) shows the visual image of an outdoor scene. Fig. 2(b) shows the registered thermal image of the same scene. The above approach was used to compute (at each pixel) the ratio between the energy conducted into the object and the absorbed radiation. The mode of this value is computed for each region [Fig. 3(a)]. The values of this internal object property occur in the expected ranges for each class of objects and help distinguish between these object classes [Fig. 3(b)]. The values are lowest for vehicles, highest for vegetation and in between for buildings and pavements. Classification of objects using this property value is discussed in [5].

2) Analyzing a Temporal Sequence of Multisensor Data: Temporal sequences of multisensor image data are commonly used for remote sensing and surveillance applications. A temporal sequence of multisensor data consisting of thermal imagery, visual imagery, and scene conditions makes possible a more reliable estimate of the imaged ob-



(a) (b)

Fig. 3. (a) Mode of the feature value computed for each region and (b) classification using the feature value.

ject's relative ability to sink/source heat radiation. Observe that the relationship between the conducted heat flux $W_{\rm cd}$ and the thermal capacitance C_T of the object is given by

$$W_{\rm cd} = C_T \frac{dT_s}{dt}. \tag{10}$$

A finite (backward) difference approximation to this equation may be used for estimating C_T as

$$C_T = W_{\rm cd} \frac{(t_2 - t_1)}{(T_s(t_2) - T_s(t_1))} \tag{11}$$

where t_1 and t_2 are the time instants at which the data were acquired, $T_s(t_1)$ and $T_s(t_2)$ are the corresponding surface temperatures, and $W_{\rm cd}$ is the conducted heat flux which is assumed to be constant during the time interval. However, $W_{\rm cd}$ does vary and an average value of $(W_{\rm cd}(t_1) + W_{\rm cd}(t_2))/2$ is used in (11).

The above method was tested on temporal sequences of thermal and visual image pairs acquired at intervals of three hours. For each thermal and visual image pair (viz. at each time instant) the method described in Section II-A was applied to evaluate the various components of surface energy exchange at each pixel. $W_{\rm cd}$ is thus evaluated for each pixel at each time instant. Equation (11) is then evaluated at each pixel, for each pair of successive data sets in the temporal sequence. The estimated values of C_T for different classes are close to those listed in Table 1, and can be used to distinguish between different classes of imaged materials. A statistically robust scheme for computing this parameter, and the experimental results obtained are described in [6].

B. Integrating Sonar and Optical Imagery

Computer analysis of underwater imagery has many important military and commercial applications. These include accurate underwater terrain assessment by autonomous underwater vehicles and remotely operated vehicles, seafloor mapping, exploration of natural resources, location of objects lost at sea, detecting mines, etc. In the past, underwater imagery has utilized two major sensors: sonar and visual. However, each approach by itself cannot provide sufficient information to constrain the imaged object's identity. For instance, in sonar imaging, surfaces which differ in roughness and other material properties may give rise to images of similar intensities. On the other hand, the optical surface reflectance of different object classes are similar and hence the visual images of their surfaces might be indistinguishable. Thus each modality needs additional information in order to uniquely determine object identity.

A physics-based approach for automated analysis of imagery acquired from underwater environments has been described in [11]. The approach described in [11] is based upon fusion of information from sonar and visual sensors to constrain interpretation of the imaged object. An overview of this approach is presented. The integrated interpretation of sonar and visual data uses a physical model, called the composite roughness model [12], which is based on the analysis of the energy transport across the interface between two fluids. This approach is used to estimate material properties of the seafloor which serve as physically meaningful features for seafloor classification.

A common use of optical imagery is for sensing the microprofile of the imaged surface [13], [14]. Standard photogrammetric techniques are applied to stereo photographs (Fig. 4) to record heights of bottom features. The Fourier transform of the autocorrelation of this relief produces the 2-D relief spectrum denoted as $W(k_x, k_y)$, where k_x, k_y are, respectively, the x and y components of the 2-D wave vector, $\mathbf{k}$. The 2-D surface relief can be modeled by

$$W(\mathbf{k}) = \beta k^{-\gamma}$$

where β and γ are dimensionless parameters used to describe the 2-D wave spectrum and k is the wavenumber, i.e., the magnitude of the 2-D wave vector $\mathbf{k}$. Hence, β and γ can be computed using the stereoscopic images of the surface.

Many studies have shown that a strong relationship exists between backscattering strength and material type. This

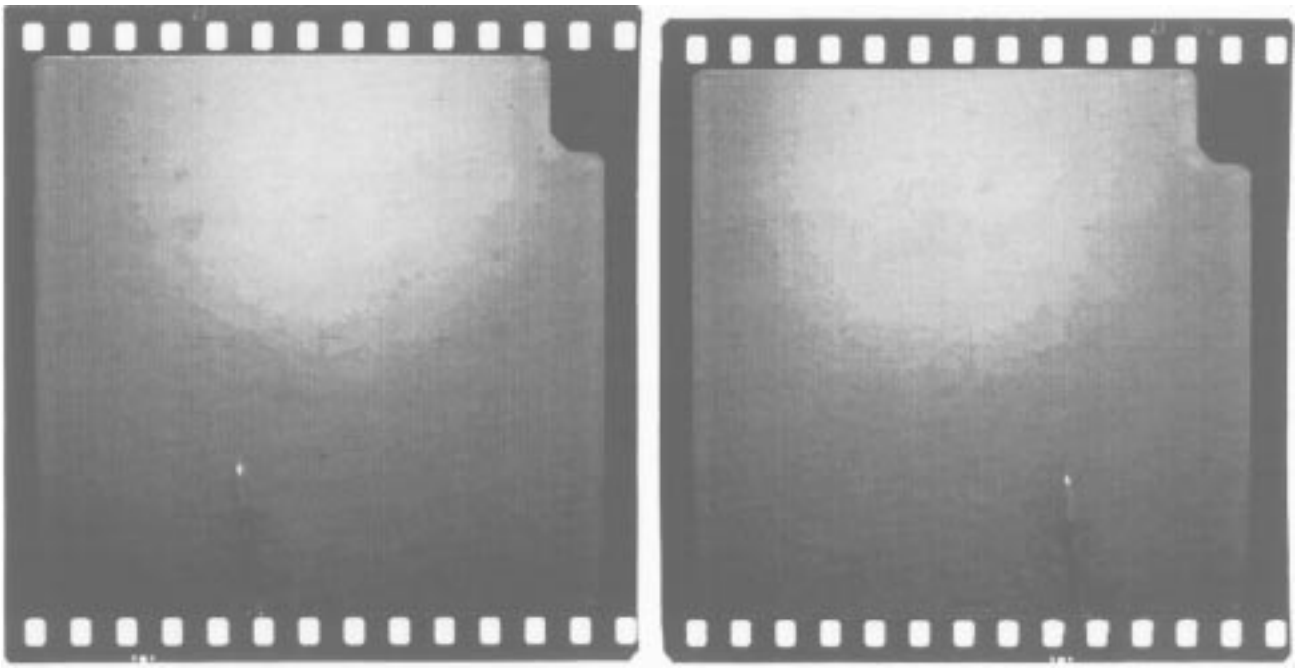


Fig. 4. Optical stereo image pair of the seafloor. Surface roughness spectrum is computed for this pair.

relationship is quantified in the composite roughness model proposed by Jackson [12]. Good agreement exists between real data and predicted values at moderate grazing angles and frequencies of 40 kHz or higher [15]. The backscattering cross section of the surface, $\sigma(\theta)$, is computed from sonar data (Fig. 5). Here, θ is the grazing angle of the acoustic energy. The composite roughness model which treats the boundary between the seafloor and water as a two-fluid interface, relates the backscattering cross section to: 1) ratio of compressional wave speed to water sound speed denoted as ν, 2) the ratio of object mass density to water mass density denoted as ρ, and 3) surface roughness.

The small-scale roughness backscattering cross section is modeled by [12]

$$\sigma_{\text{model}}(\theta) = 4k_a^4 \sin^4\theta F(\theta,\nu,\rho) W(2k_a\cos\theta, 0) \quad (12)$$

where $k_a = \omega/c$ is the acoustic wavenumber in water, W is the power spectrum of the 2-D roughness distribution obtained from analyzing the optical stereo imagery.

$$F(\theta,\nu,\rho) = \begin{cases} \frac{[(\rho-1)^2\cos^2\theta+\rho^2-\nu^{-2}]^2}{[\rho\sin\theta+(\nu^{-2}-\cos^2\theta)^{-1/2}]^4}, & \theta > \theta_c \\ \frac{[(\rho-1)^2\cos^2\theta+\rho^2-\nu^{-2}]^2}{[(1-\rho^2)\cos^2\theta+\rho^2-\nu^{-2}]^2}, & \theta < \theta_c. \end{cases} \quad (13)$$

The critical angle is given by $\theta_c = \cos^{-1}(\nu^{-1})$.

The small-scale roughness backscattering cross section is obtained from the sonar data. Thus $\sigma_{\text{data}}(\theta_i)$ is available for various values of θ_i. The power spectrum of the 2-D roughness distribution is obtained from bottom photographs. Hence, $W_i(\mathbf{k})$ is available for each of the θ_i. The backscattering cross section model, $\sigma_{\text{model}}(\theta)$, given by (12) and (13), is now fit to the data-derived backscattering cross section $\sigma_{\text{data}}(\theta_i)$. This is a nonlinear regression problem where ρ and ν are the parameters of the regression fit. A least means squared error minimization technique is adopted to compute ρ and ν.

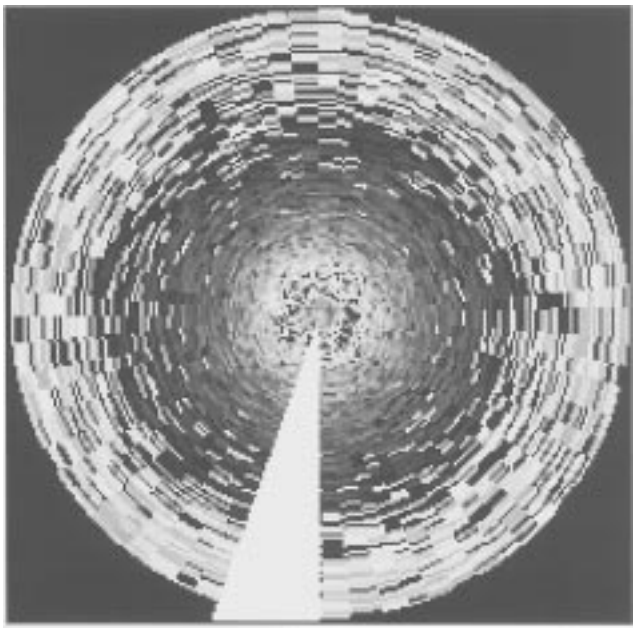


Fig. 5. Raw sonar data of seafloor. Image was formed using 69 sonar pings of a 40 kHz system. Reflections from each ping form a radial line in the image. Azimuthal resolution is 5°.

A feature vector consisting of ρ and ν is then used to classify the seafloor, e.g., as sandy sediment or sediment rich in manganese nodules using a minimum distance classifier. Table 2 lists results of analyzing imagery from a seafloor consisting of only sediment and bereft of manganese nodules.

C. Interpreting Polarized Radiation

Often, an object being imaged emanates energy (reflected or emitted energy) that has polarization information which can be "decoded" to make inferences on the type of object

Table 2 Estimated Values of ρ and ν for Real Data from Sediment. Ground Truth Values are $\rho = 1.4$, $\nu = 1.0014$. Initial Values Were $\rho_0 = 1.85, \nu_0 = 1.4$. Classification Based on Distance Values.

Ping	Estimated Values		d_{sed}	d_{Mn}	class
	ρ	ν			
0	1.07	1.0013	0.40	2.18	sed
5	1.07	1.0017	0.40	2.18	sed
10	1.08	1.0019	0.40	2.16	sed
15	1.07	1.0013	0.40	2.18	sed
20	1.07	1.0014	0.40	2.18	sed
25	1.09	1.0018	0.38	2.13	sed
30	1.08	1.0014	0.39	2.16	sed
35	1.09	1.0019	0.38	2.18	sed
40	1.08	1.0016	0.39	2.16	sed
45	1.06	1.0011	0.41	2.21	sed
50	1.07	1.0012	0.40	2.18	sed
55	1.85	1.4	0.12	0.236	sed
60	1.07	1.0012	0.40	2.18	sed

surface being sensed. Filters or sensor configurations may be used to create different images corresponding to only specific modes of polarization. Complementary information from these images yields information about object structure or material type. Two different examples are presented below—the first discusses, briefly, the analysis of optical imagery, and the second discusses the analysis of different polarization channels of SAR imagery.

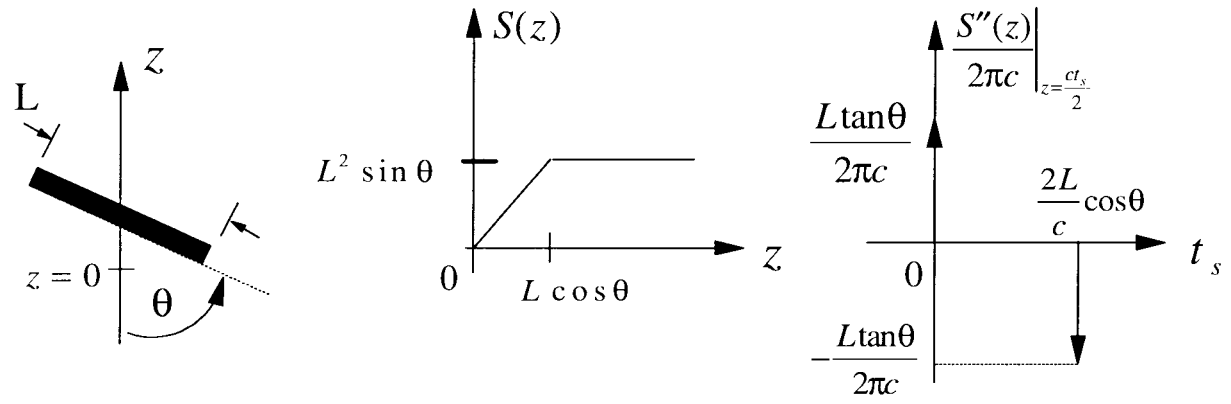
1) Polarized Optical Imagery: When light is incident on commonly occurring surfaces, the reflected light consists of either a specular component, a diffuse component or a combination of the two. If the incident light is unpolarized, the diffuse component of the reflected light remains unpolarized while the specular component of the reflected light becomes partially linear polarized. The degree to which the specularly reflected component becomes partially polarized depends on the electrical conductivity of the reflecting surface. For materials with large conductivities, the polarization is reduced. This property has been used to segment metal surfaces from dielectric surfaces [16].

Consider unpolarized light incident on a surface, with angle of incidence ψ (with respect to the surface normal), and the specularly reflected light at angle of emittance, also equaling ψ. The specular plane is defined to be the plane containing the incident ray and the reflected ray. For specularly reflected light the magnitude of the polarization component perpendicular to the specular plane is larger than magnitude of the polarization component parallel to the specular plane. The Fresnel reflection coefficients, $F_\perp$ and $F_\parallel$, corresponding to the polarization components that are, respectively, perpendicular to and parallel to the specular plane have values between zero and one. The magnitude ratio of these polarization components is the polarization Fresnel ratio (PFR) and equals $F_\perp/F_\parallel$. The value of PFR varies with the specular angle of incidence, ψ. Wolff argues that for values of ψ between 30° and 80° the value of PFR is below 2.0 for metals and above 2.0 for dielectrics. Hence, the PFR can be used to differentiate these two classes of metals. Different thresholds on PFR may be used to detect metals coated with translucent insulating materials.

One way of computing the PFR of an object in the scene is to acquire a sequence of images using a camera in front of which is mounted a polarizer that is incrementally rotated. The maximum intensity, $I_{\max}$, at a pixel location, and the minimum intensity, $I_{\min}$, at that location are noted. The ratio $I_{\max}/I_{\min}$ approximates the PFR if the diffuse component of the reflection is much weaker than the specular component—which holds for many surfaces, but is violated for a surface with highly diffuse albedo, such as a sheet of bond paper. An alternative method of estimating the PFR is to acquire a smaller number of images corresponding to fewer orientations of the polarizer. The intensity variation at any pixel is then modeled as a sinusoidal function—and the parameters of the fitted model are used to find $I_{\max}$ and $I_{\min}$.

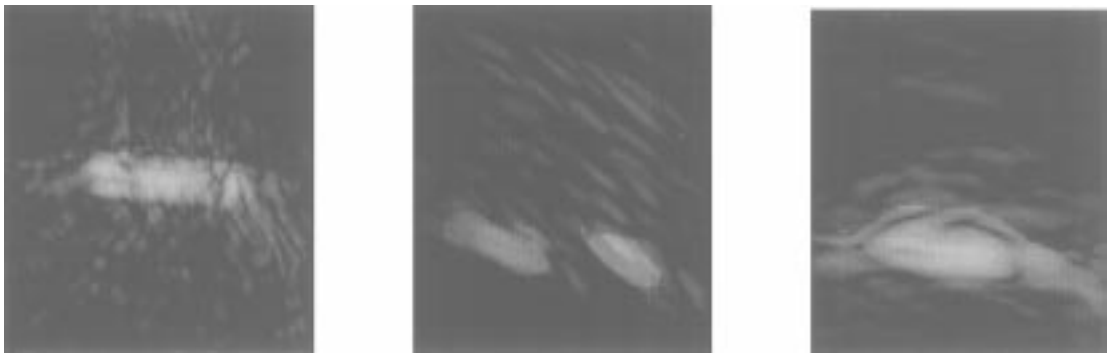
The above approach is hampered when the illumination is a distributed sources instead of a point source, and also then the specular angle of incidence is close to the grazing angle. More recently, the approach has been extended to distinguish edges caused by specularities from those caused by occluding boundaries [17]. Further extensions have been reported to include extended light sources, and object recognition applications [18].

2) Polarized SAR Imagery: A physics-based approach has been reported for the interpretation of ultra-wideband (UWB) SAR imagery for object classification [19], [20]. The UWB sensor (50 MHz to 1 GHz) is used for detection of man-made objects, including those which are obscured by a random media, e.g., vehicles obscured by foliage, and buried objects. An electromagnetic model is used to predict the backscatter received from the scene objects. The backscatter from a scene object consists of two components: 1) an "early time" portion that precedes a 2) late-time portion. The early-time (physical optics) portion of the backscatter is due to reflected energy and is highly dependent on the structure of the object surface that is illuminated by the incident energy. The late time response is due to resonant modes excited by the incident energy, and is dependent on the gross shape of the object. Hence, the early time response from scene objects that differ in structure, such as vehicles and trees, have different sensitivity to aspect angle of the incident wave. Aspect angle sensitivity of the backscatter is extracted by reconstructing the SAR image of an object over smaller subapertures of the full synthetic aperture. This multi-aperture approach provides important angular information while still maintaining detectable levels in the subsequent SAR images. The backscatter from man-made objects differs from natural objects both in angular and polarimetric dependence. This is primarily due to the fact that manmade objects are more planar in structure and more conductive than natural objects. The structure of natural objects, such as trees and rocks, changes in a random manner with viewpoint, and hence, the aspect-angle dependence of backscatter from natural objects is more random that that for man-made objects. This characteristic is used to develop feature vectors using a multi-aperture approach to discriminate vehicles hidden in a deciduous forest from surrounding objects of little interest.



(a) (b) (c)

Fig. 6. Approximation of early-time portion of backscatter for a flat plate: (a) the incident wave illuminates the plate at aspect angle θ, (b) the silhouette area function $S(z)$ is determined, and (c) the second derivative of the function is used in the physical optics approximation to estimate the backscatter.



(a) (b) (c)

Fig. 7. UWB-SAR image of a pickup truck (a) using full aperture, (b) quarter subaperture which is off-broadside of truck, and (c) quarter subaperture which is along broadside of truck.

Considering the physical optics or early time portion of the energy scattered by the imaged surface, and assuming an impulse incident wave, the resulting scattered wave is given by

$$\vec{H}(\vec{r},t) = \frac{1}{2\pi c}\frac{d^2 S(t_s)}{dt_s^2} \tag{14}$$

where the $\vec{H}$ is the contribution of the incident wave in generating the local surface current at $\vec{r}$, $S(t_s)$ is the silhouette area of the scatterer projected along the direction of propagation of the incident wave, z, and $t_s = \frac{2z}{c}$; $z = 0$ at the beginning of the scatterer. Equation (14) can be used to provide a simple approximation of the scattered field due to the incident impulse wave [21] and it is evident that the scattered field will be dependent on aspect angle since $S(t_s)$ changes with aspect angle to the scatterer. An example of this calculation is shown in Fig. 6 for a flat plate of length L where the incident wave arrives at aspect angle θ to the surface of the plate. We observe that the backscatter will consist of two finite valued peaks which are due to surface discontinuities at the edges and tips of the plate. These closely spaced peaks have a maximum magnitude when the direction of propagation for the incident wave is normal to the plate surface.

The aspect angle sensitivity of an imaged object can be computed as follows. First, a SAR image is reconstructed using all N range profiles that are collected at various aspect angles across the synthetic aperture angle interval, $\Delta\theta_a$. An object of interest is detected when a pixel in the image is found to exceed a local contrast threshold, i.e., a "glint" is detected. Next, the N range image profiles are grouped into subsets of n_p $(n_p < N)$ consecutively acquired (adjacent) range profiles, each subset comprising a single subaperture. An image of an object of interest can be reconstructed using the range profiles of a single subaperture. When this is done for each subaperture, aspect angle sensitivity of the object's backscatter can be determined by examining the variation of the object's backscatter across the different subapertures.

The SAR system used in the study was fully polarimetric, in that the radar has the capability of transmitting two orthogonal polarizations consecutively and receiving two orthogonal polarizations simultaneously at each point along the aperture. Therefore, multi-aperture processing can be applied to both co- and cross-polarized channels. Given two orthogonal polarization bases X and Y, $\alpha_{xx}(\theta)$ and $\alpha_{yy}(\theta)$ are defined to be the backscatter energy as a function of subaperture position (aspect-angle) for the co-polarized

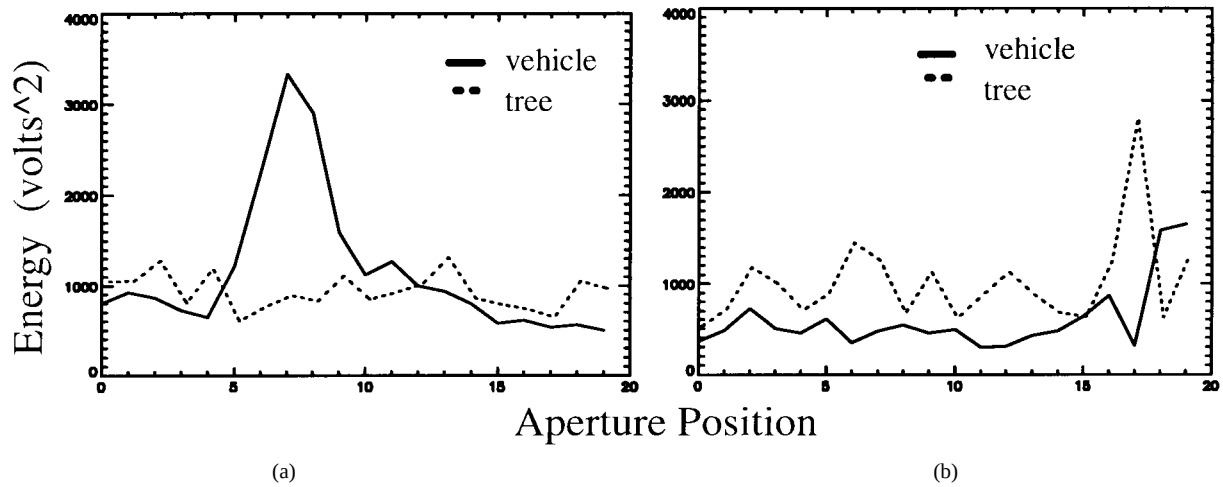


Fig. 8. Aspect angle dependence of backscatter from (a) broadside vehicle (truck) and clutter (tree) where "specular flash" is observed at aperture position seven and (b) truck whose specular flash is not observed and tree which appears aspect angle dependent.

channels where first and second subscripts indicate transmit and receive polarizations, respectively. Similarly, multi-aperture processing of cross polarized channels results in $\alpha_{xy}(\theta)$ and $\alpha_{yx}(\theta)$—however, due to symmetry, cross polarized returns are generally similar to one another. A normalized correlation coefficient is defined

$$r_{xx\_yy} = \frac{\int_{\Delta\theta_a} \alpha_{xx}(\theta)\alpha_{yy}(\theta)\,d\theta}{\left[\int_{\Delta\theta_a} \alpha_{xx}^2(\theta)\,d\theta \cdot \int_{\Delta\theta_a} \alpha_{yy}^2(\theta)\,d\theta\right]^{1/2}}. \quad (15)$$

The coefficients r_{xx_xy} and r_{yy_xy} may be defined in a similar manner. The coefficient value is bounded by $0 \leq |r_{xx\_yy}| \leq 1$, $r_{xx\_yy} = 0$ if and only if α_{xx} and α_{yy} are independent, and $r_{xx\_yy} = 1$ if and only if $\alpha_{xx} = K\alpha_{yy}$, where K is a constant. A feature vector can now be formulated whose elements are correlation coefficients r_{xx_yy}, r_{xx_xy}, r_{yy_xy}.

Consider man-made vehicles and natural objects in an imaged scene. Vehicles are generally metallic and have more large, flat planar surfaces than objects occurring in nature such as trees, rocks, etc. Hence, at certain aspect angles, the radar may observe a large increase in backscatter from vehicles due to specular reflections. In contrast, trees, tree limbs, and foliage will not exhibit this large increase in backscatter, and will have more random fluctuations with aspect angle due to small radar cross section and random distribution in a resolution cell. The backscatter is independently random for both co- and cross-polarized channels. Therefore, the correlation coefficient defined in (15) will be low for natural objects. The response for vehicles will be similar for both co- and cross-polarized channels and, consequently, the correlation coefficient defined in (15) will be larger for vehicles than for natural objects.

The UWB SAR images of a pickup truck using the full aperture and two subapertures are shown in Fig. 7. The aspect-angle signatures for the broadside truck and a tree are shown in Fig. 8(a). An increase in backscatter energy is observed near aperture position seven, where the specular flash from the vehicle is observed. The tree, on the other hand, has a random but bounded backscatter energy across the aperture, as expected. However, two factors may degrade detection performance or increase false alarm rates using this approach: 1) due to changing foliage density with aspect angle, clutter can often appear to be aspect angle dependent, and 2) the specular flash from the vehicle may not be observed along the synthetic aperture, making it difficult to discriminate between vehicles and clutter [Fig. 8(b)]. These difficulties are overcome by using the feature vector described above, which combines both angular and polarimetric diversity information (Fig. 9).

D. Other Imaging Sensors

The physics-based approach has been used for the integration of other types of sensors, in addition to the imaging modalities discussed above. A brief overview of some examples are given below.

1) Color Image Interpretation: Color may also be considered to be multisensory information since irradiation in three different spectral bands are sensed. Initial work in physics based interpretation of color imagery was conducted by Klinker *et al.* [22] who discuss the segmentation of objects using physical models of color image generation. Their model consists of a dichromatic reflection model that is a linear combination of surface reflection (highlights) and reflection from the surface body. The combined spectral distribution of matte and highlight points forms a skewed T-shaped cluster in red-green-blue space, where the matte points lie along one limb of the T and the highlight points lie along the other limb. Principal component analysis of color distributions in small nonoverlapping windows provides initial hypotheses of the reflection type. Adjacent windows are merged if the color clusters have similar orientations. These form "linear hypotheses." Next, skewed T-shaped clusters are detected. This specifies the dichromatic model used to locally resegment the color image via a recursive region merging process. Thus a combination of

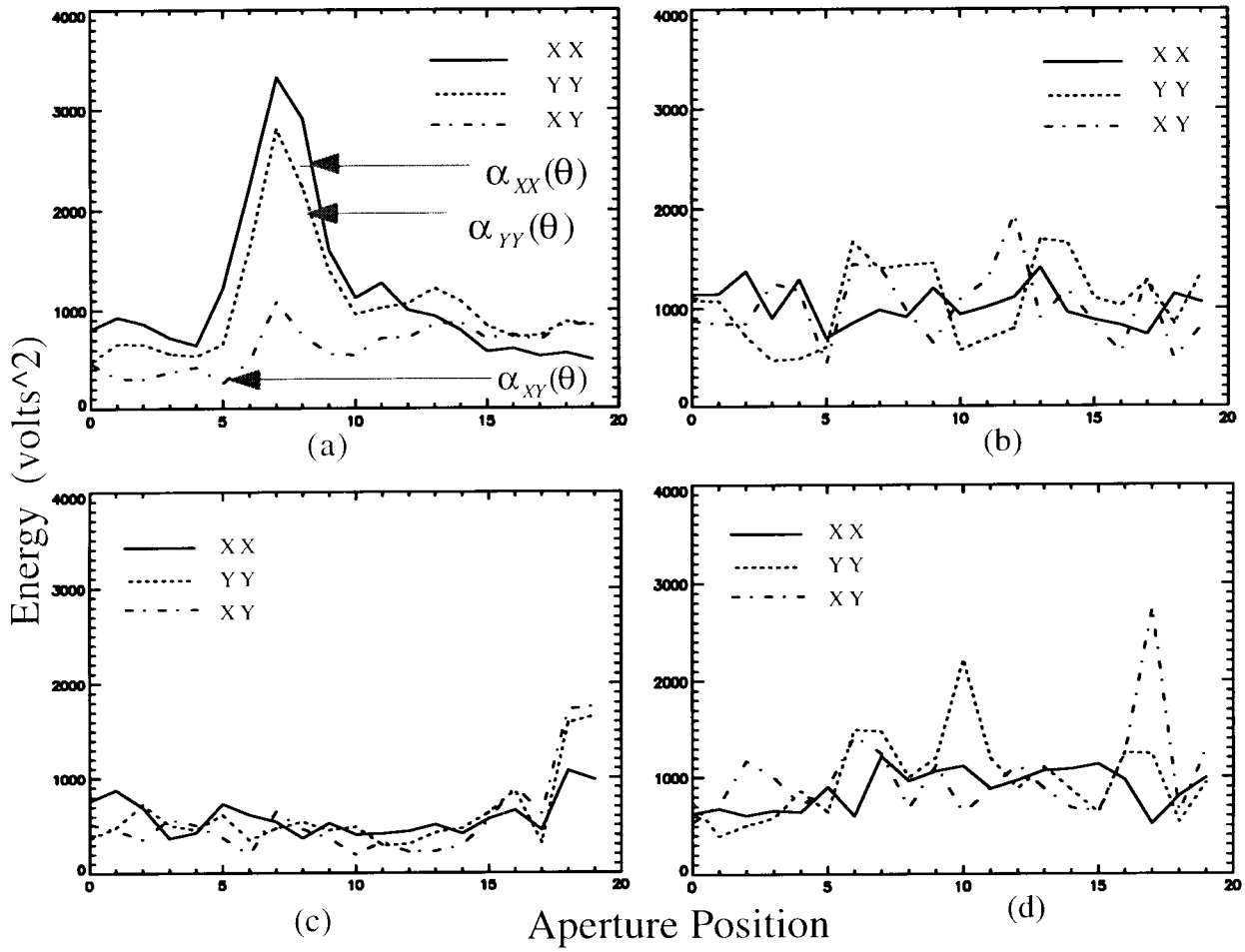


Fig. 9. Full polarimetric multi-aperture response of (a) pickup truck oriented broadside with specular flash at position seven, producing a feature vector $R = (0.93\ 0.57\ 0.25)$, (b) tree $(R = (0.01\ 0.04{-}0.04))$, (c) pickup truck whose specular flash is not observed $(R = (0.75\ 0.71\ 0.92))$, and (d) tree which appears aspect angle dependent $(R = (0.36\ 0.23{-}0.32))$.

bottom-up and top-down processing segments images into regions corresponding to objects of different color.

Healey [23], [24] reports a color segmentation approach that uses a reflection model that includes metallic as well as dichromatic surfaces. The segmentation algorithm considers the color information at each pixel to comprise a Gaussian random vector with three variables. Segmentation is achieved by a recursive subdivision of the image and by the analysis of resulting region level statistics of the random vector. This physics based approach has been extended for object recognition using color information [25], [26].

2) Radar and Optical Imagery: The integration of visual images and low-resolution microwave radar scattering cross sections to reconstruct the three-dimensional (3-D) shapes of objects for space robotic applications is discussed in [27]. Their objective is to "combine the interpreted output of these sensors into a consistent world view that is in some way better than its component interpretations." The visual image yields contours and a partial surface shape description for the viewed object. The radar system provides an estimate of the range and a set of polarized radar scattering cross sections, which is a vector of four components. An "intelligent decision module" uses the information derived from the visual image to find a standard geometrical shape for the imaged object. If this is possible, then a closed form expression is used to predict the radar cross section. Otherwise, an electromagnetic model uses the sparse surface description to compute the radar cross section using a finite approximation technique. The unknown shape characteristics of the surface are then solved for iteratively based on minimizing the difference between the predicted and sensed radar cross section. This technique is illustrated by a simulation reported in [27].

III. Integration of Image and Model Information

Model-based object recognition is a commonly adopted paradigm in computer vision and pattern recognition, wherein features extracted from a region in the image are compared with those stored in a model—resulting in classification of the image region. This paradigm is easily extended to include a physics-based approach where the features stored in object models are based on material properties. A further extension is possible in which the model stores the values of physical properties (such

as density, specific heat, conductivity) of the materials that comprise the object. For an image region under consideration, a hypothesis is made of the identity of the object, and hence of the materials present in the scene. The values of these materials' properties as stored in the model, along with image gray levels at the locations at which these materials are hypothesized to occur, are used to compute a feature. The value of this feature is used to verify or refute the hypothesis. The feature is based on a physics based approach similar to that described in Section II-A. An overview of such an approach used for recognition of objects in long wave infrared (LWIR) imagery is presented below [28]–[30].

A. Features Using Model and Image Information

Applying the principle of conservation of energy at the surface of an imaged object yielded the energy balance equation (3) and (5) in Section II-A. While those equations applied for thin plate objects, a more general form that accounts for nonuniform temperature distributions in the material is given by

$$W_{\text{abs}} = W_{\text{rad}} + W_{\text{cv}} + W_{st} + W_{\text{cnd}} \tag{16}$$

where W_{abs}, W_{rad}, and W_{cv} are as defined earlier, W_{st} is the energy used to raise the temperature of the local infinitesimal volume at the surface of the object, and W_{cnd} is the energy conducted into the interior of the object. Hence we have $W_{st} = C_T \frac{dT_s}{dt}$ and $W_{\text{cnd}} = k\frac{dT_s}{dx}$, where k is the thermal conductivity, and x is the depth from the surface.

The new energy balance equation may be rewritten in the following linear form

$$a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 = \bar{\mathbf{a}}^T\bar{\mathbf{x}} = 0. \tag{17}$$

Using the expressions for the various energy components as presented in the previous section we can express each term in the above expression as

$$\begin{array}{ll} a_1 = C_T & x_1 = -\dfrac{dT_s}{dt} \\ a_2 = k & x_2 = -\dfrac{dT_s}{dx} \\ a_3 = -(T_s - T_{\text{amb}}) & x_3 = h \\ a_4 = -\sigma(T_s^4 - T_{\text{amb}}^4) & x_4 = \epsilon \\ a_5 = \cos\theta_I & x_5 = W_I\alpha_s. \end{array} \tag{18}$$

Note that a calibrated LWIR image provides radiometric temperature (assuming $\epsilon = 0.9$ which is true for most surfaces). Hence a_3 and a_4 can be computed from the LWIR image alone (and knowledge of the ambient temperature), while a_1, a_2, and a_5 are provided by the model when the identity and pose of the object is hypothesized. The "driving conditions," or unknown scene parameters that can change from scene to scene are given by the x_i, $\quad i = 1, \cdots, 5$. Thus each pixel in the thermal image equation (18) defines a point in 5-D thermophysical space.

Consider two different LWIR images of a scene obtained under different scene conditions and from different viewpoints. For a given object, N points are selected such that 1) the points are visible in both views and 2) each point lies on a different component of the object which differs in material composition and/or surface orientation. Assume (for the nonce) that the object pose for each view, and point correspondence between the two views are available (or hypothesized). A point in each view yields a measurement vector $\bar{\mathbf{a}} = (a_1, a_2, a_3, a_4, a_5)^T$ with components defined by (18) and a corresponding driving conditions vector $\bar{\mathbf{x}} = (x_1, x_2, x_3, x_4, x_5)^T$. Let a collection of N of these vectors compose a $5 \times N$ matrix, $\mathbf{A} = (\bar{\mathbf{a}}_\mathbf{1}, \bar{\mathbf{a}}_\mathbf{2}, \cdots, \bar{\mathbf{a}}_\mathbf{N})$ for the first scene/image. These same points in the second scene will define vectors that compose a $5 \times N$ matrix, $\mathbf{A}' = (\bar{\mathbf{a}}'_\mathbf{1}, \bar{\mathbf{a}}'_\mathbf{2}, \cdots, \bar{\mathbf{a}}'_\mathbf{N})$. The driving condition matrix, $\mathbf{X} = (\bar{\mathbf{x}}_\mathbf{1}, \bar{\mathbf{x}}_\mathbf{2}, \cdots, \bar{\mathbf{x}}_\mathbf{N})^\mathbf{T}$, from the first scene and $\mathbf{X}' = (\bar{\mathbf{x}}'_\mathbf{1}, \bar{\mathbf{x}}'_\mathbf{2}, \cdots, \bar{\mathbf{x}}'_\mathbf{N})^\mathbf{T}$ from the second scene, are each of size $N \times 5$.

Since the N points are selected to be on different material types and/or different surface orientations, the thermophysical diversity causes the vectors $\bar{\mathbf{a}}_\mathbf{1}, \bar{\mathbf{a}}_\mathbf{2}, \cdots, \bar{\mathbf{a}}_\mathbf{N}$ to span $\Re^5$, as will also the vectors, $\bar{\mathbf{a}}'_\mathbf{1}, \bar{\mathbf{a}}'_\mathbf{2}, \cdots, \bar{\mathbf{a}}'_\mathbf{N}$. Without loss of generality, assume that vectors $\bar{\mathbf{a}}_\mathbf{1}, \cdots, \bar{\mathbf{a}}_\mathbf{5}$ span $\Re^5$ and also that $\bar{\mathbf{a}}'_\mathbf{1}, \cdots, \bar{\mathbf{a}}'_\mathbf{5}$ span $\Re^5$. These five points specify the 5×5 measurement matrices, $\mathbf{A} = (\bar{\mathbf{a}}'_\mathbf{1}, \bar{\mathbf{a}}'_\mathbf{2}, \cdots, \bar{\mathbf{a}}'_\mathbf{5})$ and $\mathbf{A}' = (\bar{\mathbf{a}}'_\mathbf{1}, \bar{\mathbf{a}}'_\mathbf{2}, \cdots, \bar{\mathbf{a}}'_\mathbf{5})$, corresponding to the first and second scene, respectively. The point selection process here is analogous to the selection of characteristic 3-D points in the construction of geometric invariants. Since $\mathbf{A}$ and $\mathbf{A}'$ are of full rank, there exists a linear transformation S such that $\mathbf{A} = S\mathbf{A}'$. Hence, an induced nonsingular linear transformation can be shown to exist between $\bar{\mathbf{x}}_\mathbf{k}$ and $\bar{\mathbf{x}}'_\mathbf{k}$.

Consider the measurement vector $\bar{\mathbf{a}}$ of a point as defined in (18). From one scene to the next we expect the two object properties—thermal capacitance C_T and conductance k—to remain constant. Hence the linear transformation $M : \bar{\mathbf{a}} \rightarrow \bar{\mathbf{a}}'$ must be of the form

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} \end{bmatrix}. \tag{19}$$

The transformation of a measurement vector from one scene to the next is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} \end{bmatrix} \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ a_{3,j} \\ a_{4,j} \\ a_{5,j} \end{bmatrix} = \begin{bmatrix} a'_{1,j} \\ a'_{2,j} \\ a'_{3,j} \\ a'_{4,j} \\ a'_{5,j} \end{bmatrix} \quad j = 1, 2, \cdots, 5. \tag{20}$$

The first two elements, the thermal capacitance and the thermal conductance, are held constant and the other scene dependent elements are allowed to change. In general, we have $M\mathbf{A} = \mathbf{A}'$, where $\mathbf{A}$ and $\mathbf{A}'$ are the 5×5 matrices derived from the two scenes and for the chosen points.

Algebraic elimination of the transformation parameters using four copies of the linear form (17) subjected to

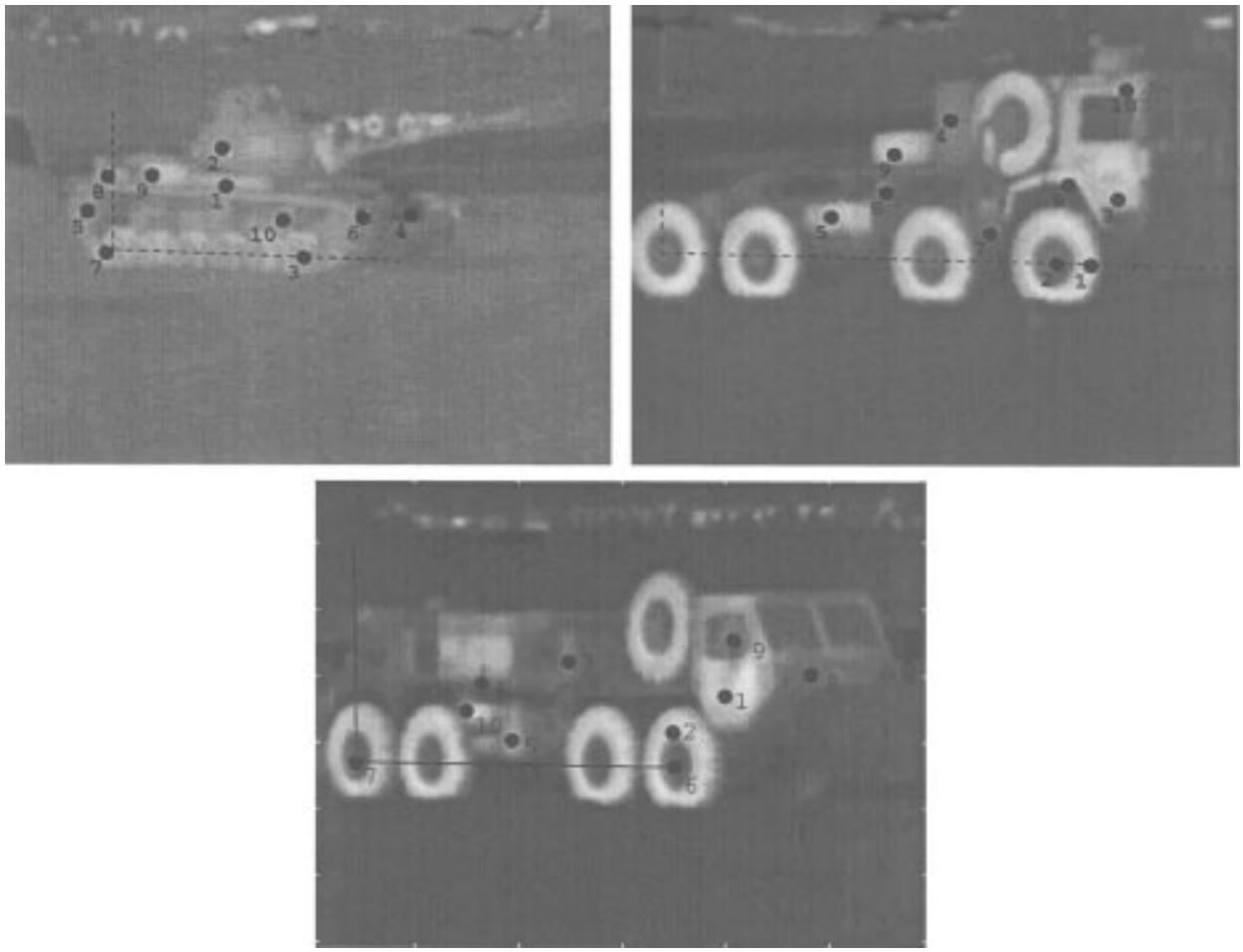


Fig. 10. Three of the five vehicles used to test the object recognition approach. From top left (clockwise): tank, truck 1, and truck 2. The axis superimposed on the image show the object centered reference frames. The numbered points indicate the object surfaces used to form the measurement matrices. These points are selected such that there are a variety of different materials and/or surface normals within the set.

the transformation (19) provides us with invariants, i.e., functions of the measurement vectors for a scene, which remain constant and are independent of the driving conditions and the transformation. This elimination may be performed by using recently reported symbolic techniques [31]. The five invariant functions derived by this elimination process can be divided into two types. Each is a ratio of determinants. The first type of invariant function uses determinants formed from components of three of the four vectors.

$$I1 = \frac{\begin{vmatrix} a_{1,3} & a_{2,3} & a_{4,3} \\ a_{1,4} & a_{2,4} & a_{4,4} \\ a_{1,5} & a_{2,5} & a_{4,5} \end{vmatrix}}{\begin{vmatrix} a_{2,3} & a_{3,3} & a_{4,3} \\ a_{2,4} & a_{3,4} & a_{4,4} \\ a_{2,5} & a_{3,5} & a_{4,5} \end{vmatrix}} \tag{21}$$

where $a_{i,j}$ is a jth component of the ith vector (ith point).

The second type is formed from components of all four vectors.

$$I2 = \frac{\begin{vmatrix} a_{1,1} & a_{2,1} & a_{3,1} & a_{4,1} \\ a_{1,2} & a_{2,2} & a_{3,2} & a_{4,2} \\ a_{1,4} & a_{2,4} & a_{3,4} & a_{4,4} \\ a_{1,5} & a_{2,5} & a_{3,5} & a_{4,5} \end{vmatrix}}{\begin{vmatrix} a_{1,1} & a_{2,1} & a_{3,1} & a_{4,1} \\ a_{1,3} & a_{2,3} & a_{3,3} & a_{4,3} \\ a_{1,4} & a_{2,4} & a_{3,4} & a_{4,4} \\ a_{1,5} & a_{2,5} & a_{3,5} & a_{4,5} \end{vmatrix}} \tag{22}$$

where $a_{i,j}$ is a jth component of the ith vector (ith point). Since the four measurement vectors span $\Re^4$, we can assume without loss of generality that the denominator determinants in (21) and (22) are nonzero. The first type has $\binom{4}{3} = 4$, independent functions given four points and second type has one.

In order for the invariant feature to be useful for object recognition not only must the values of the feature, be invariant to scene conditions but the value must be different if the measurement vector is obtained from a scene that does not contain the hypothesized object, and/or if the hypothe-

Table 3 Values of the I1–Type Feature Used to Identify the Vehicle Class, Truck 1. The Feature Consisted of Point Set {4,7,8,10} Corresponding to the Points Labeled in the Figure. The Feature Value is Formed Using the Thermophysical Model of Truck 1 and the Data from the Respective Other Vehicles. When this Feature is Applied to the Correctly Hypothesized Data of the Tank, It has a Mean Value of -0.57 and a Standard Deviation of 0.13. This I1–Type Feature Produces a Good Stability Measure of 4.5, and Good Separability Between Correct and Incorrect Hypotheses. The Feature Values for Incorrect Hypotheses are at Least 3.32 Standard Deviations Away from the Mean Value for the Correct Hypothesis

Hypothesis: Data From:	Truck 1 Truck 1	Truck 1 Tank	Truck 1 Van	Truck 1 Car	Truck 1 Truck 2
11 am	-0.70	27.28	0.33	$-\infty$	0.68
12 pm	-0.71	0.09	4.83	15.58	4.15
1 pm	-0.45	0.68	0.00	11.73	4.6e12
2 pm	-0.66	-1.00	$-\infty$	71.23	-1.00
3 pm	-0.40	-1.00	$-\infty$	-1.00	-1.00
4 pm	-0.54	$-\infty$	∞	5.42	22.29
9 am	-0.68	1.38	-1.00	$-6.66e14$	-7.03
10 am	-0.45	-1.00	$-\infty$	6.50	∞

sized pose is incorrect. Since the formulation above takes into account only feature invariance but not separability, a search for the best set of points that both identifies the object and separates the classes must be conducted over a given set of points identified on the object.

B. Scheme for Object Recognition

The hypothesize-and-verify scheme for object recognition consists of the following steps: 1) extract geometric features, e.g., lines and conics, 2) for image region, r, hypothesize object class, k, and pose using, for example, geometric invariants as proposed by Forsyth *et al*. [32], 3) use the model of object k and project visible points labeled $i = 1, 2, \cdots$ onto image region r using scaled orthographic projection, 4) for point labeled i in the image region, assign thermophysical properties of point labeled i in the model of object k, 5) using gray levels at each point and the assigned thermophysical properties, compute the measurement matrices A and A', and hence compute the feature $f^k(r)$ using (14) or (15), and finally, 6) compare feature $f^k(r)$ with model prototype $\hat{f}_k$ to verify the hypothesis.

The above technique was applied to real LWIR imagery acquired at different times of the day. Several types of vehicles were imaged. Three of these vehicles are shown in Fig. 10. Several points were selected (as indicated in the figures) on the surfaces of different materials and/or orientation. Given an image of a vehicle, 1) the pose of the vehicle is assumed known, then 2) the front and rear wheels are used to establish an object centered reference frame. The center of the rear wheel is used as the origin, and center of the front wheel is used to specify the direction and scaling of the axes. The coordinates of the selected points are expressed in terms of this 2-D object centered frame. For example, when a truck-1 vehicle is hypothesized for an image actually obtained of a tank or some unknown vehicle, the material properties of the truck-1 are used, but image measurements are obtained from the image of the tank at locations given by transforming the coordinates of the truck-1 points (in the truck-1 centered coordinate frame) to the image frame computed for the unknown vehicle. The features are computed based on the image data and model information. Table 3 shows interclass and intraclass variation for a feature of type $I1$ under the truck-1 hypothesis—for images obtained from different vehicles at eight different times over two days. The feature of type $I2$ exhibits similar behavior—high between-class separation and low within-class variation.

IV. Synthesis of Multisensory Imagery

Model derived features and imagery are useful in the development of recognition systems that use trainable classifiers, or other model-driven approaches. In order to achieve low error rates it is necessary to train the classifier on multiple images of the same object in a variety of scene conditions. This requires the existence of a large training database. Creating this database using real scene imagery is expensive and sometimes impossible. It is also difficult to maintain a large database for training. Furthermore, if the classifier is to be trained on multimodal data the size of the data set is even greater. An attractive solution to the storage and data acquisition problems is to create accurate artificial images of the objects. Not only are storage space and database maintenance requirements lessened, but computer generation of object imagery provides greater flexibility in specifying environmental and object conditions. A unified model of different modes of sensing, such as visual, thermal, and ladar, augments the ability to quickly generate a large multisensory data set for training, without the expense of field work. These reasons have motivated the development of synthetic image generation and feature prediction systems.

One of the principal issues facing researchers in the area of recognition systems is the establishment of models that can simulate the image-generating physical processes peculiar to each of the different imaging modalities. Further, it is attractive to use a single modeling scheme that can be used for the different modalities. Early work in this area by Oh *et al*. [33] addressed the generation of thermal and visual imagery using 3-D object models stored in the form of octrees [34]. The simulation of energy exchange and energy flows within the object gave rise to predictions of surface temperatures and hence the thermal image, while the 3-D

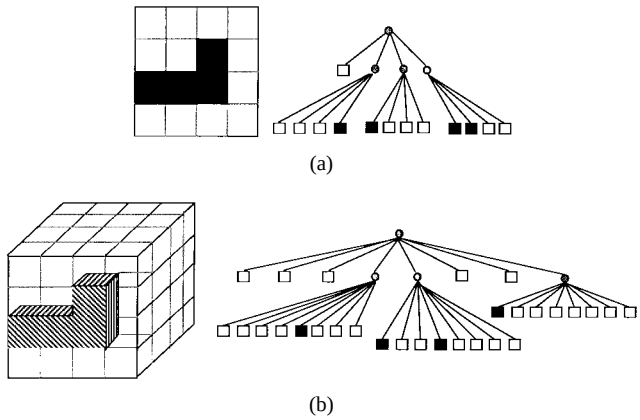


Fig. 11. The quadtree representation of a binary image, and the octree representation of a 3-D object.

surface was used to create the visual image. Further work by Karthik *et al.* [35], increased the accuracy of the model by incorporating nonhomogeneities in the object model. This approach was further augmented to allow the simulation of laser radar (range, reflectance, and doppler) imagery in addition to the visual and infrared modalities [36], [37]. An overview of this approach is presented below.

When the goal of the simulation is to generate prediction of only feature values used for recognition—as opposed to imagery—then, a more efficient scheme may be used. Rather than model a fully tessellated 3-D object, one can represent (at a coarse level of resolution) only the major components of an object, each consisting of uniform material properties. Thus very few nodes are used in the simulation of energy flows and exchange. Accurate prediction of multisensory physics-based features have been produced by this approach [38].

A. 3-D Object Model Construction

The 3-D object model is represented in the form of an octree. The octree of a 3-D object is obtained by subdividing the universe into eight equal octants successively until each octant corresponds either to an empty space or to a volume inside the object. The octree is an extension of a quadtree which may be used to represent 2-D objects. The quadtree of a binary image is obtained by subdividing the image into four quadrants successively until each quadrant is either entirely black or white. Each quadrant is then a node in the tree. Each node has either four children or is a leaf node. The nodes can be white, black or gray. A gray node is not a leaf node and has four child nodes that can be gray, white, or black nodes. The quadtree and octree representations are illustrated in Fig. 11.

Multiple silhouettes are used as the input to the octree formation. The contour of each silhouette is smoothed using a tension spline and the contour normal at each contour pixel is computed. For each view, a region contour (RC) quadtree is constructed, wherein a node can be either black (object), white (nonobject), contour (contour pixel), or gray (parent of others). The contour normal is stored with each contour node. Using the RC quadtree from different views,

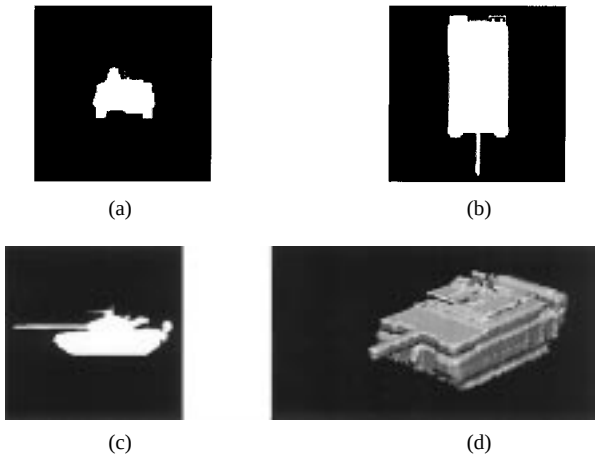


Fig. 12. Visual Image of object reconstructed from three silhouettes showing false surfaces resulting from the "basic" volume intersection approach.

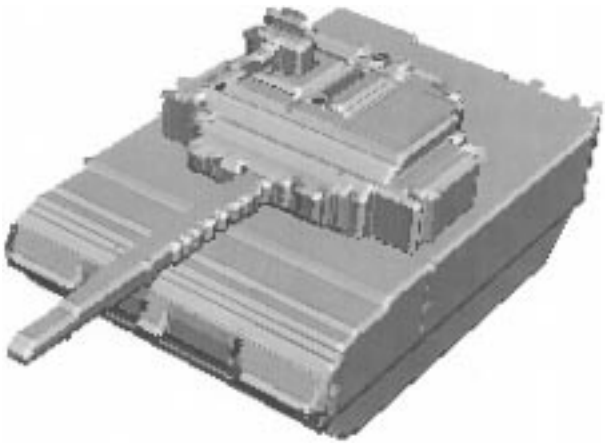


Fig. 13. Visual image of object reconstructed using the improved volume intersection technique based on silhouette partitioning.

a volume intersection is performed and a volume surface (VS)-octree is constructed, wherein the nodes are black (internal), white (empty), surface, or gray (parents of the others). Each surface node is encoded with the surface normal. The above approach suffers from the disadvantage that concave boundaries in the silhouettes result in false volumes in the 3-D reconstruction. However, these errors may be avoided by using line-drawing information to partition the silhouettes into different regions prior to volume intersection [37].

B. Generating Different Imagery

1) Visual Image Generation: It is easy to generate the visual image of an object from its octree representation. All the surface nodes are scanned and checked for visibility. Only the visible faces of the leaf nodes (voxels) are projected to form the visual image. Lambertian or other reflectance functions may be assumed. Either orthographic, weak perspective, or perspective projection can be employed. Any arbitrary viewpoint as well as the direction of illumination can be specified. Figs. 12 and 13 are examples of results generated by this process.

2) Thermal Image Generation: Each voxel (leaf node) in the octree is encoded with thermophysical material properties and links are attached to spatially adjacent nodes. Applying the conservation of energy to each node we have the energy balance law

$$\frac{\rho c v\left(T_n^{k+1}-T_n^k\right)}{dt} = \alpha G_s \cos\theta A_s + \dot{q}v - \sum_{i=1}^{nadj} \frac{kA_{ni}\left(T_n^{k+1}-T_i^{k+1}\right)}{d_{ni}} - hA_s\left(T_n^{k+1}-T_{\text{amb}}\right) - h_r A_s\left(T_n^{k+1}-T_{\text{amb}}\right) \quad (23)$$

where ρ is the mass density, c is the specific heat, v is the volume of the node, T denotes the temperature of the node, subscript n denotes the node number (spatial location), superscript $k+1$ represents the time instant, dt denotes the time interval, G_s is the magnitude of solar irradiation, θ is the angle between the direction of solar irradiation and the surface normal, α is the absorptivity of the node, A_s is the surface area of the node (relevant face), $\dot{q}$ is the volumetric heat generation, $nadj$ is the number of adjacent nodes, k is the thermal conductivity, A_{ni} is the area of the interface between the node n and each of the adjacent nodes, d_{ni} is the distance between the centroid of node n and the centroid of the ith adjacent node, h is the convection coefficient, T_{amb} is the ambient temperature, and h_r is the radiation coefficient. The above process is simulated for the desired environmental and object conditions to predict surface temperatures.

The gray level, L_T, produced by an infrared camera operating in the 8 μm–12 μm band and at distances of up to a few hundred meters from commonly occurring objects is related to the surface temperature, T_s, by

$$K_a L_T + K_b = 0.9 \int_{\lambda_1}^{\lambda_2} \frac{C_1}{\lambda^5(\exp(C_2/\lambda T_s)-1)} d\lambda \quad (24)$$

where $\lambda_1 = 8$ μm, $\lambda_2 = 12$ μm, C_1 and C_2 are constants with $C_1 = 3.742 \times 10^8$ W μm/m^2 and $C_2 = 1.439 \times 10^4$ μm K. K_a and K_b are constants based on the camera and digitizing parameters. Thus the surface temperatures can be projected using the same technique as for visual imagery and then transformed by the above expression to produce the thermal image.

3) Ladar Range Image: An AM ladar range image is created by sensing the difference between the phase of the modulation envelope of the transmitted signal and that of the received signal. Since the phase difference is periodic with a period of 2π, the range measurement is only available as a fraction of the wavelength. The ambiguity interval for AM ladar can be written as

$$R_{\text{am}} = \frac{c}{2f_m} \quad (25)$$

where R_{am} is the ambiguity interval, c is the speed of light, and f_m is the modulation frequency.

A range measurement is subject to noise. Let the actual range be given by

$$R = nR_{\text{am}} + R_\epsilon \quad (26)$$

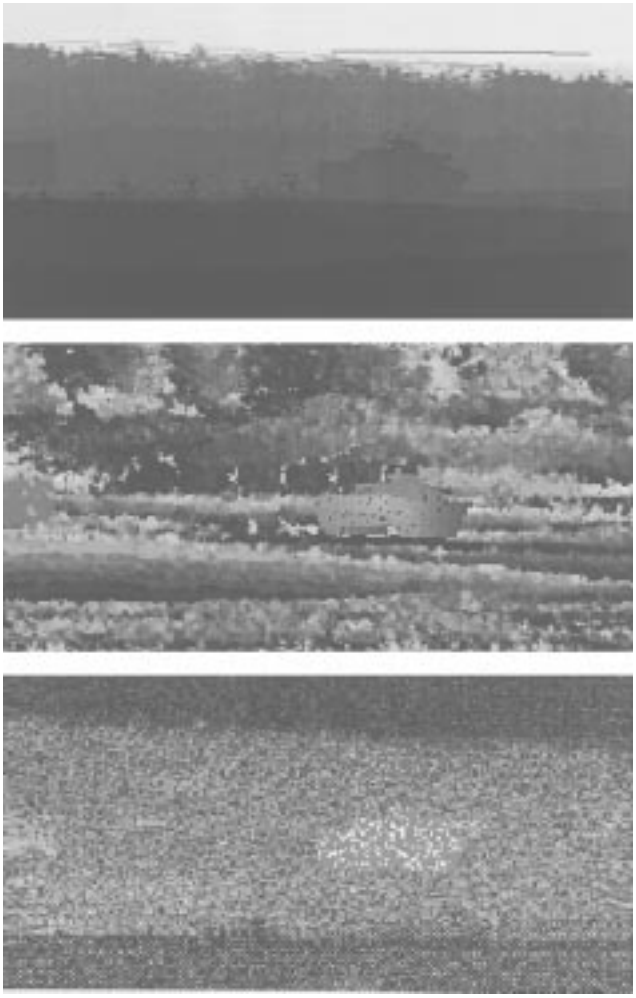


Fig. 14. Simulated ladar images of a T-72M1 tank model synthesized from silhouettes. The images contain real backgrounds with the simulated object placed in the scene. The top image is the absolute range image. The middle image is the fine range image with a ambiguity interval of 19.8 m. The bottom image is the reflectance image. Constructive and destructive interference noise has been simulated on the object and its effect on the reflectance image is shown. Destructive interference prevents extraction of range information. This effect on the fine range image is shown.

where n is the number of ambiguity intervals, $R_\epsilon \in [0, \lambda/2]$ is the fraction of the ambiguity interval measured by the sensor, and λ is the wavelength of the modulating envelope. The range including effects of noise is

$$R' = nR_{\text{am}} + R_\epsilon + R_\eta \quad (27)$$

$R_\epsilon + R_\eta \in [0, \lambda/2]$ is the output of the laser radar imaging system. $R_\eta \in \left[-\frac{\lambda}{4}, \frac{\lambda}{4}\right]$ is the noise [39]

$$R_\eta = \frac{c}{2\pi f_m} \arcsin \frac{a_n(t)\sin(2\pi f_m t + P_m(t))}{\sqrt{\text{SNR}}a(t)} \quad (28)$$

where $a_n(t)$ is the Rayleigh distributed noise amplitude, $a(t)$ is the Rayleigh distributed signal amplitude, SNR is the detector signal to noise ratio, and $P_m(t)$ is the uniformly distributed phase error. Typically, a modulating frequency wavelength of 15–20 m is used for relative range, depending on the object to be sensed. In addition to noise in the range values, speckle noise due to rough surfaces is

common. Appropriate statistical models of this process can be used to simulate speckle noise in the reflectance image. Fig. 14 shows an example of ladar image synthesis.

V. Conclusion

In this paper we described a new class of techniques for integrating information from different sensing modalities—a physics-based approach that used appropriate physical models of the image generation process. These models usually rely on the principle of conservation of energy, which when applied to the imaged scene provides analytical constraints between material properties and the imaged gray levels. This approach makes available physically meaningful object features that are highly specific—they provide good separation between different classes of objects.

The advantages of multisensory approaches to computer vision are evident from the discussions in the previous sections. The integration of multiple sensors and/or multiple sensing modalities is an effective method of minimizing the ambiguities inherent in interpreting perceived scenes. Although the multisensory approach is useful for a variety of tasks including pose determination, surface reconstruction, object recognition, and motion computation, among others—the current paper addressed the problem of object recognition. Several problems that were previously difficult or even impossible to solve because of the ill-posed nature of the formulations are converted to well-posed problems with the adoption of a multisensory approach.

Recent and continuing developments in multisensory vision research may be attributable to several factors, including: 1) new sensor technology that makes affordable previously unexplored sensing modalities, 2) new scientific contributions in computational approaches to sensor fusion, and 3) new insights into the electro-physiological mechanisms of multisensory perception in biological perceptual systems. Most of the progress to date may be attributed to the second cause listed above. The development of new, affordable sensors is currently an important and active area of research and may be expected to have a significant future impact on the capabilities of vision systems. For example, the availability of low cost imaging laser ranging sensors, passive infrared sensors, and high frequency radar imagers would provide significant impetus to research in developing multisensor-based autonomous navigation, object recognition, and surface reconstruction techniques. Many lessons from nature are yet to be learned neuro-physiological and psycho-physiological studies of natural perceptual systems. Such studies may provide useful clues for deciding what combination of sensing modalities are useful for a specific task, and may also provide new computational models for intersensory perception.

References

[1] N. Nandhakumar and J. K. Aggarwal, "Multisensory computer vision," *Advances in Computers*, vol. 34, M. C. Yovits, Ed. New York: Academic, 1992, pp. 59–111.

[2] D. Hilbert, *Theory of Algebraic Invariants*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[3] N. Nandhakumar and J. K. Aggarwal, "Integrated analysis of thermal and visual images for scene interpretation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 10, pp. 469–481, July 1988.

[4] ——, "Multisensor integration—experiments in integrating thermal and visual sensors," in *Proc. IEEE Computer Soc. 1st Int. Conf. on Computer Vision*, London, U.K., June 1987, pp. 83–92.

[5] ——, "Thermal and visual information fusion for outdoor scene perception," in *Proc. IEEE Int. Conf. on Robot. Automation*, Philadelphia, PA, Apr. 1988, pp. 1306–1308.

[6] N. Nandhakumar, "Robust physics-based sensor fusion," *J. Opt. Soc. Amer., JOSA-A*, special issue on Physics-Based Computer Vision, vol. 11, no. 11, pp. 1–9, Nov. 1994.

[7] P. H. Hartline, L. Kass, and M. S. Loop, "Merging of modalities in the optic tectum: Infrared and visual information integration in rattlesnakes," *Sci.*, vol. 199, pp. 1225–1229, 1978.

[8] E. A. Newman and P. H. Hartline, "The infrared 'vision' of snakes," *Scientif. Amer.*, vol. 246, no. 3, pp. 116–127, Mar. 1982.

[9] P. H. Hartline, "The optic tectum of reptiles: Neurophysiological studies," *Comparative Neurology of the Optic Tectum*, H. Vanegas, Ed. New York: Plenum, 1984, pp. 601–618.

[10] F. P. Incropera and D. P. DeWitt, *Fundamentals of Heat Transfer*. New York: Wiley, 1981.

[11] N. Nandhakumar and S. Malik, "Multisensor integration for underwater scene classification," *Appl. Intell.*, vol. 5, pp. 207–216, 1995.

[12] D. R. Jackson, D. P. Winebrenner, and A. Ishimaru, "Application of the composite roughness model to high-frequency bottom backscattering," *J. Acoust. Soc. Amer.*, vol. 79, no. 5, pp. 1410–1422, May 1986.

[13] D. M. Owen, "A multi-shot stereoscopic camera for close-up ocean-bottom photography," in *Deep-Sea Photography*, J. B. Hersey, Ed. Baltimore: Johns Hopkins, 1967, pp. 95–105.

[14] C. J. Shipek, "Deep-sea photography in support of underwater acoustic research," in *Deep-Sea Photography*, J. B. Hersey, Ed. Baltimore: Johns Hopkins, 1967, pp. 89–94.

[15] S. Stanic *et al*., "High-frequency acoustic backscattering from a coarse shell ocean bottom," *J. Acoust. Soc. Amer.*, vol. 85, no. 1, pp. 125–136, Jan. 1989.

[16] L. B. Wolff, "Polarization-based material classification from specular reflection," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 10, pp. 1059–1071, Nov. 1990.

[17] ——, "Scene understanding from propogation and consistency of polarization-based constraints," in *Proc. IEEE Computer Soc. Computer Vision and Patt. Recognition Conf.*, Seattle, WA, June 21–23, 1994, pp. 1000–1005.

[18] ——, "Reflectance modeling for object recognition and detection in outdoor scenes," in *Proc. ARPA Image Understanding Workshop*, Palm Springs, CA, Feb. 12–15, 1996, pp. 799–803.

[19] R. Kapoor and N. Nandhakumar, "A physics-based approach for detecting man-made objects in UWB-SAR imagery," *IEEE Computer Soc. Workshop on Physics Based Models for Computer Vision*, Cambridge, MA, June 18–20, 1995, pp. 33–39.

[20] ——, "Features for detecting obscured objects in ultra-wideband (UWB) SAR imagery using a phenomenological approach," to be published in *Patt. Recognition*.

[21] E. M. Kennaugh and D. L. Moffatt, "Transient and impulse response approximations," in *Proc. IEEE*, vol. PROC-53, pp. 893–901, 1965.

[22] G. J. Klinker, S. A. Shafer, and T. Kanade, "Image segmentation and reflection analysis through color," in *Proc. DARPA Image Understand. Workshop*, Cambridge, MA, 1988, pp. 838–853.

[23] G. Healey, "Using color to segment images of 3-D scenes," in *Proc. SPIE Conf. Applicat. AI*, Orlando, FL, Apr. 1991, vol. 1468, pp. 814–825.

[24] G. Healey, S. Shafer, and L. Wolff, Eds., *Physics Based Vision: Principles and Practice, COLOR*. Boston: Jones and Bartlett, 1992.

[25] G. Healey and D. Slater, "Using illumination invariant color histogram descriptors for recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recognition*, Seattle, WA, June 21–23, 1994, pp. 355–360.

[26] M. Swain and D. Ballard, "Color indexing," *Int. J. Computer Vision*, vol. 7, pp. 11–32, 1991.

[27] S. W. Shaw, R. J. P. deFigueiredo, and K. Kumar, "Fusion of radar and optical sensors for space robotic vision," in

Proc. IEEE Robot. Automat. Conf., Philadelphia, PA, 1988, pp. 1842–1846.
[28] N. Nandhakumar, V. Velten, and J. Michel, "Thermophysical affine invariants from IR imagery for object recognition," *IEEE Computer Soc. Workshop on Physics Based Models for Computer Vision*, Cambridge, MA, June 18–20, 1995, pp. 48–54.
[29] J. D. Michel, T. Saxena, N. Nandhakumar, and D. Kapur, "Using elimination methods to compute thermophysical algberaic invariants from infrared imagery," in *Proc. AAAI-96*, Seattle WA, 1996.
[30] N. Nandhakumar *et al*., "Robust thermophysics-based interpretation of radiometrically uncalibrated IR images for ATR and site change detection," *IEEE Trans. Image Process.*, Dec. 1996.
[31] D. Kapur, Y. N. Lakshman, and T. Saxena, "Computing invariants using elimination methods," in *Proc. IEEE Int. Symp. on Computer Vision*, Coral Gables, FL, Nov. 21–23, 1995, pp. 97–102.
[32] D. Forsyth *et al*., "Invariant descriptors for 3D object recognition and pose," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 13, Oct. 1991.
[33] C. Oh, N. Nandhakumar, and J. K. Aggarwal, "Integrated modeling of thermal and visual image generation," *IEEE Conf. on Computer Vision and Patt. Recognition*, San Diego, June 4–8, 1989.
[34] C. H. Chien and J. K. Aggarwal, "Volume/surface octrees for the representation of 3D objects," *Computer Vision, Graphics and Image Process.*, vol. 36, pp. 100–113, 1986.
[35] S. Karthik, N. Nandhakumar, and J. K. Aggarwal, "Non-homogeneous 3-D objects for thermal and visual image synthesis," *SPIE Conf. on Applicat. AI*, Orlando, FL, 1991.
[36] J. D. Michel and N. Nandhakumar, "Unified octree-based object models for multisensor fusion," *2nd IEEE Workshop on CAD Based Vision*, Champion, PA, Feb. 8–11, 1994, pp. 181–168.
[37] ____, "Unified 3D models for multisensor image synthesis," *Computer Vision, Graphics, and Image Process.*, vol. 57, no. 4, July 1995.
[38] G. Hoekstra and N. Nandhakumar, "Quasiinvariant behavior of thermophysical features for interpretation of multisensory imagery," *Opt. Engin.*, vol. 35, no. 3, pp. 1–14, Mar. 1996.
[39] J. Leonard and E. G. Zelnio, "Intrinsic separability of laser radar imagery," in *Proc. 2nd Automatic Target Recognizer Syst. and Technol. Conf.*, Mar. 17, 1992.

N. Nandhakumar (Senior Member, IEEE) received the B.E. (honors) degree in electronics and communication engineering from the P.S.G. College of Technology, University of Madras, India, in 1981, the M.S.E. degree in computer, information and control engineering from the University of Michigan, Ann Arbor, in 1983, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 1987.

He is currently a Senior Member of the Technical Staff at Electroglas, Inc., Santa Clara, CA, where he is leading the development of new machine vision technology for semiconductor wafer inspection. He has also served as Assistant Professor of Electrical Engineering at the University of Virginia, and Director of the Machine Vision Laboratory from 1989 to 1996. The results of his research have appeared in more than 70 papers in journals, conference proceedings, and book chapters. He is Associate Editor of *Pattern Recognition*, and has been the Guest Editor of several special issues on sensor fusion.

Dr. Nandhakumar was awarded the 1993 Young Faculty Teaching Award by University of Virginia's Electrical Engineering Department. He has been Chairman of the 1994, 1993, and 1992 SPIE Conferences on Sensor Fusion and Aerospace Applications, Co-Chairman of the 1991 SPIE Conference on Applications of Artificial Intelligence, member of the international advisory committee of the 2nd Asian Conference on Computer Vision, and member of the program committees of the 1993 and 1997 IEEE CVPR Conferences, and the 1st and 2nd IEEE International Conferences on Image Processing. He is a member of the SPIE.

J. K. Aggarwal (Fellow, IEEE) is the Cullen Professor of Electrical and Computer Engineering and Director of the Computer and Vision Research Center at The University of Texas at Austin, where he has served on the faculty since 1964. His research interests include computer vision, parallel processing of images, and pattern recognition. He is the author or editor of seven books and 31 book chapters and author of over 160 journal papers, as well as numerous proceedings papers and technical reports.

Dr. Aggarwal was the recipient of the 1996 Technical Achievement Award of the IEEE Computer Society.