Year I Final Technical Report

Submitted to

Office of Naval Research

Multi Sensor Information Integration and Automatic Understanding

Contract Number: N00014-05-C0294

Submitted by

Signal Innovations Group
1009 Slater Road
Suite 200
Durham, NC 27703

22 August 2006

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
22 AUG 2006	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Multi Sensor Information Integration and Automatic Understanding	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Signal Innovations Group 1009 Slater Road Suite 200 Durham, NC 27703**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**28**	

Table of Contents

Executive Summary

This program addresses *Automatic Image Understanding* and *Automatic Integration of Disparate Sources of Information*. The research is being pursued collaboratively by the Signal Innovations Group, Inc. (SIG), Lockheed Martin Missiles & Fire Control (LMM&FC), and the NAVAIR Weapons Division at China Lake. The techniques have a rigorous mathematical foundation, buttressed by many years of ONR basic (6.1) research pursued at Duke University, NAVAIR, and LMM&FC. The techniques are particularly focused on asymmetric warfare, urban warfare, guerrilla warfare, and port/base security, for which automatic integration of disparate sources is particularly important, typically with very limited if any *a priori* training data.

Concerning **automatic image understanding**, we are principally considering image sequences (video). The approaches utilize the new field of semi-supervised learning. Specifically, most existing Automatic Target Recognition (ATR) approaches are supervised, in the sense that they require an *a priori* training set of labeled data $\mathbf{D}_L$. The set $\mathbf{D}_L$ is composed of example signatures (features) and their associated identity (label). These data are typically employed to design a classifier, with the hope that the labeled training set $\mathbf{D}_L$ is well matched statistically to the unlabeled data $\mathbf{D}_U$ to which the ATR algorithm is applied. Such supervised algorithms are vitiated by the inherent differences in training and testing data ($\mathbf{D}_L$ and $\mathbf{D}_U$, respectively) found in practice. In addition, in conventional techniques the classifier is applied to each element of $\mathbf{D}_U$, one at a time, without accounting for the cumulative contextual information inherent to $\mathbf{D}_U$. The *semi-supervised algorithms* employed here ameliorate the limitations of conventional approaches by performing learning based on all available data, both labeled and unlabeled. By explicitly employing $\mathbf{D}_U$ in design of the classifier, the algorithm automatically accounts for context and for changing sensing conditions.

The performance of the semi-supervised classifier is directly related to the features extracted from the imagery and video. LMM&FC has done extensive research and testing on target detection algorithms based upon Quadratic Correlation Filtering (QCF) theory which project the image data onto a subspace which is optimal for discriminating targets versus clutter. The merger of these features into the *semi-supervised* classification algorithms offers the possibility of higher detection and classification rates, less training requirements, and adaptive algorithm updates. These techniques and features are also being employed within the rigorous joint classification and feature optimization (JCFO) algorithm developed at Duke and now transitioned to SIG. The JCFO is a state-of-the-art data-mining framework that accounts for the importance of given features/data for classification, thereby merging the feature selection and classifier-design stages.

The semi-supervised classifiers utilize Bayesian kernel algorithms, employing graph-theoretic techniques. Graph-based techniques account for all available data, both labeled and unlabeled (respectively, $\mathbf{D}_L$ and $\mathbf{D}_U$), in a mathematically rigorous framework. For video, we are exploiting the spatial information from a given image as well as the temporal information associated with a changing scene. Time-varying spatial information is well characterized via a hidden Markov model (HMM), for which a given scene is

represented by (hidden) states. The states may reflect the (unobserved) intentions of an individual being monitored via video, *e.g.*, an asymmetric threat in an urban environment. Almost all previous research with HMMs has been performed in the context of supervised algorithms, for which an *a priori* labeled training set $\mathbf{D}_L$ is assumed and the context inherent to $\mathbf{D}_U$ is not exploited. For the problem of interest here the HMM has been placed within a semi-supervised framework. The HMM will therefore be adaptable for interpretation of video, under the realistic scenario for which *a priori* training data may be quite limited (as anticipated for asymmetric threats).

For the thrust on **automatic integration of disparate sources of information**, we are developing algorithms that autonomously integrate and manage an arbitrary number and range of sensors. Within the information-integration framework, the algorithms merge sensor data with information that may be provided by personnel in theatre (human intelligence, or HUMINT). The integration of multiple and disparate information sources is manifested via a graph-based statistical prior, and the sensor management is implemented by computing the *expected* information gain associated with particular actions. The management of resources (sensors and personnel) is manifested by accounting for costs associated with particular actions (*e.g.*, time, bandwidth and/or energy requirements).

Our knowledge of a given environment is defined by the available labeled and unlabeled data, $\mathbf{D}_L$ and $\mathbf{D}_U$, respectively. We augment $\mathbf{D}_L$ and $\mathbf{D}_U$ to improve our knowledge base, by deploying sensors and personnel. For example, personnel or near-range sensors may be employed to acquire labels for particular elements in $\mathbf{D}_U$, with the subsequent labeled examples used to augment $\mathbf{D}_L$. Alternatively, one may employ new sensors to enhance our understanding of $\mathbf{D}_U$, without the costly task of acquiring labels. Since the information content of such actions depends on what is observed (what labels or data are collected), and these observations are unavailable in advance, we must compute the *expected* information gain. In the research being pursued this is implemented in a mathematically rigorous framework, utilizing the *theory of optimal experiments*, where here an "experiment" consists of deploying sensors and/or personnel and acquiring new data about the environment. Actions are prioritized based on the balance between *expected* information gain and deployment costs. For imagery this framework will be implemented using state-of-the-art Bayesian kernel algorithms, and for video the time variation will be exploited via an HMM formulation. The HMM framework is also being utilized when performing multi-aspect imaging of a given scene. These techniques define the particular disparate viewing angles which are optimal for the classification performance. This is the premise behind the LMM&FC *collaborative ATR algorithms* which have been researched under recent ONR sponsorship.

The investigators at SIG, LMM&FC and NAVAIR have a long history of successful collaborations that is being leveraged in this program. NAVAIR is funded by ONR separately.

I. Concept of Operation for the Navy

The future of modern warfare lies in managing and exploiting netted and distributed systems with algorithms for automatic integration of disparate sources of information coupled with automated image understanding (or automatic target recognition – ATR). The hardware infrastructure for such systems is advancing, and is being further revolutionized by the FORCEnet program, for which the program is targeted. As an example, the pictorial in Fig. 1 depicts an artist's concept of the littoral warfare environment. It shows the key elements of a netted and distributed system such as sensors, platforms, communication links, and weapons. The same technology may be used in an urban setting, for detection of asymmetric threats, utilizing techniques that learn typical behavior, with anomalous events appearing with low likelihiood.

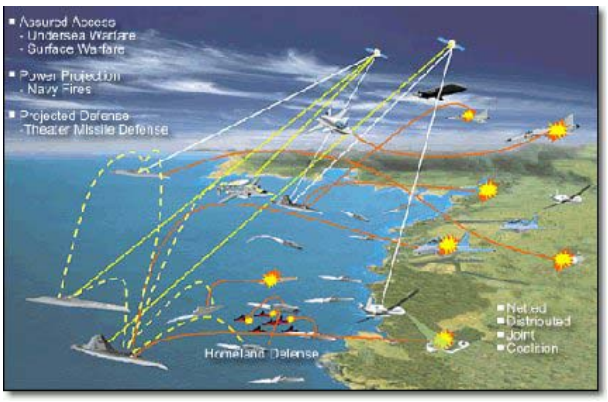

Figure 1: A conceptual netted and distributed system for Littoral Warfare

The magnitude of data implied by such a distributed sensor network such as in Fig. 1 is enormous. It requires robust algorithms for automatic image understanding and ATR which utilizes and tasks all the available sensors in the network in a way which is optimal for the exploitation of information. Given that multiple sensors of disparate information are available, we seek to take advantage of the network and its resources to address information integration and sensor management within the context of asymmetric warfare, urban warfare, guerrilla warfare and port/base security. An environment such as that in Fig. 1 is of interest for port security, with similar issues of interest for the other targeted applications (asymmetric threats in urban and suburban settings). The program is developing *general* algorithms of interest to asymmetric threats, with specific examples and demonstrations (milestones) dictated by ONR priorities and by NAVAIR/LMM&FC experimental resources (which are substantial).

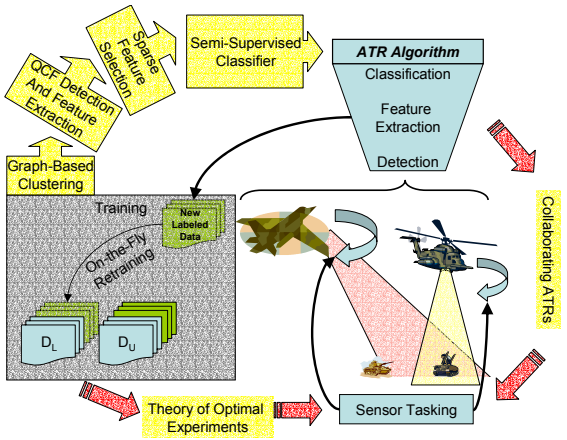

Figure 2. An ATR suite (detection, feature extraction, and classification) is designed from partially labeled data. Results are used to tasks sensors, analyze features, and update training.

In the past, the approach to automatic target detection/recognition (ATD/R) has been open loop and treated as a stand-alone problem. Thus, the sensing of information has not been guided by the methodology

for processing it. Defining algorithms for image and video understanding, taken in conjunction with the integration of disparate sources of information is key to the data exploitation. Also, one must take care to define, with a sound mathematical framework, algorithms which: process the data locally, adapt to heterogeneous environments, use all available data, and work in conjunction with the sensors.

The SIG, LMM&FC, and NAVAIR team is developing and testing algorithms that consider the whole reconnaissance/surveillance system, where individual algorithm components are clearly synergistic with other algorithms, sensors, and the overall mission. Fig. 2 outlines a notional data processing environment where components to be researched have a clear relation to other components and the overall mission.

The close coupling of the research with NAVAIR will allow demonstrations on measured prototype multi-sensor data, with the particular test examples to be investigated defined in collaboration with ONR (we anticipate deploying video sensors at China Lake in the next year of funding). The quantitative specifications for how the products will improve operational performance include:

- Demonstrated reduction in the amount of labeled training required data for algorithm design, with this achieved via semi-supervised classifiers that incorporate context. Specifically, the semi-supervised classifiers integrate information from labeled "training" data *and* unlabeled "testing" data. We will demonstrate a quantitative reduction in the required amount of training data, as well as quantitative improvements in detection performance (*e.g.*, ROC curves) when environmental conditions are changing.

- Demonstration of feature-based image and video information extraction, whereby joint classification and feature optimization (JCFO) is achieved. This manifests a direct mapping from analog sensor data to information (A/I, rather than conventional A/D). This will be quantified in terms of the number of bits required for conventional digital compression (A/D) *vis-à-vis* the number of bits required for analog to information (A/I). The classification performance of A/D and A/I algorithms will also be quantified (*e.g.*, ROC curves, confusion matrices).

- Within the context of sensor management and information integration, we will quantify the number of optimally chosen sensor and HUMINT deployments required to achieve a given objective, *vis-à-vis* non-optimal and essentially random sensor deployment. We will also quantify the performance of optimal sensor deployment *vis-à-vis* fixed or random sensor deployment (*e.g.*, ROC curves).

II. Project Schedule and Milestones

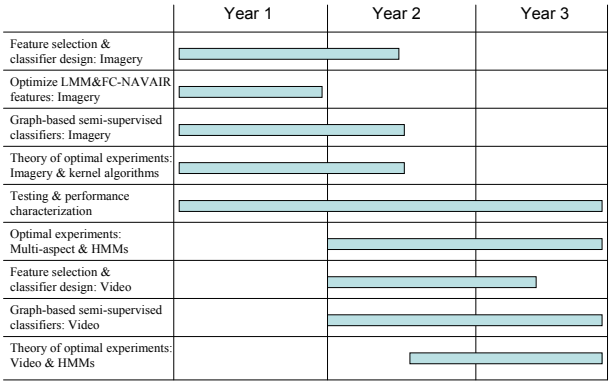

Table 1. Summary of the schedule for the tasks listed in Sec. II.

During the first year of the program we have focused principally on information sources involving video. As detailed below, we have developed techniques that remove an arbitrary background, for extraction of general foreground (moving) objects. The algorithms track arbitrary moving objects, and by maintaining shape information the algorithms seamlessly mitigate occlusions. In the next year of funding the extracted moving object will be analyzed via statistical algorithms, and the sensor-management algorithms will be used to control the camera(s) characteristics. Below we summarize the Year I milestones and associated progress.

Year 1 Milestones: Using video, demonstrate automatic extraction of features from imagery. This will be demonstrated for the realistic scenario in which the number of labeled examples is far smaller than the number of unlabeled examples to be classified.

All Phase I milestones have been met, and have been demonstrated to the sponsor. In addition, SIG has provided ONR and China Lake the underlying Matlab^{TM} code used to implement the algorithms. The algorithms provide real-time processing of video at 30 frames per second, with automatic learning of the background statistics, automatic extraction of moving objects, estimation of the shape of all moving entities, and Bayesian techniques for handling occlusions. The algorithms have been demonstrated on several different complex data sets.

In Year II we will deploy cameras at China Lake for further testing of the algorithms, on video of interest to the Navy (e.g., for base security). We will develop the HMM statistical models for learning typical behavior, and we will develop cost-sensitive sensor-management agents for optimal control of the multiple sensors.

III. Overall Technical Approach and Deliverables

III.1 Overarching Approach

The program tasks involve the following technical challenges:

- Development of classifiers that utilize all available data, both labeled and unlabeled. This manifests a *semi-supervised algorithm* that naturally adapts as new unlabeled data are acquired. In addition, when classifying any given item, the algorithm must utilize information from all labeled and unlabeled data (*i.e.*, the classification must be *placed in the context of all information sources*). This framework must also integrate different classes of information (multiple sensors and personnel).
- Development of algorithms that *autonomously integrate and manage multiple information sources* (sensors and personnel) as a given scene is interrogated. Collection of new information is here termed an "experiment", and the efficient utilization of multiple resources will be achieved using the *theory of optimal experiments*.
- Development of algorithms that perform *joint feature selection and classifier design*, such that only the most relevant components of an information source are extracted. Such feature extraction must be realized in a multi-sensor setting, integrating both labeled and unlabeled data.
- The above challenges must be addressed in the context of new sensor modalities, with *video* being an important target. The multiple types of data (images, video, waveforms, and human intelligence) must be integrated within a single framework.

In the following we provide details on how these challenges will be addressed within the research program.

III.1.1 Semi-supervised classifier

Assume for simplicity we have a binary problem, where label y=1 corresponds to targets of interest and label y=0 corresponds to false targets (clutter). The similarity of any two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined in terms of a generally non-linear kernel $K(\mathbf{x}_i,\mathbf{x}_j)$, where $K(\mathbf{x}_i,\mathbf{x}_j)$ is large when $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in feature space. For an arbitrary feature vector $\mathbf{x}$ the classifier seeks to learn a function

$$f(\mathbf{x}|\mathbf{w}) = \sum_{n=1}^{N_L} w_n K(\mathbf{x},\mathbf{x}_n) + w_o = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) \qquad (1)$$

where $\boldsymbol{\varphi}(\mathbf{x}) = \{1,\ K(\mathbf{x},\mathbf{x}_1),\ K(\mathbf{x},\mathbf{x}_2),\ ...,K(\mathbf{x},\mathbf{x}_{N_L})\}^T$ and $\mathbf{w} = \{w_o,\ w_1,\ w_2,\ ...,w_{N_L}\}^T$. The probability that a given feature vector $\mathbf{x}$ is associated with y=1 or y=0 is expressed as

$$p(y|\mathbf{x},\mathbf{w}) = \left[\exp[f(\mathbf{x}|\mathbf{w})]/\{1+\exp[f(\mathbf{x}|\mathbf{w})]\}\right]^y \left[1/\{1+\exp[f(\mathbf{x}|\mathbf{w})]\}\right]^{1-y}$$

This formulation is readily extended to an arbitrary number of label types (beyond the binary y=0/1 problem discussed here for simplicity). The probability of the label given feature vector $\mathbf{x}$ is now expressed as

$$\mathrm{p}(y|\mathbf{x}) = \int \mathrm{p}(y|\mathbf{x},\mathbf{w})\mathrm{p}(\mathbf{w}|\mathbf{x})\mathrm{d}\mathbf{w}$$

where $\mathrm{p}(\mathbf{w}|\mathbf{x})$ is a prior on the weights $\mathbf{w}$. Previously researchers have imposed a sparseness prior on $\mathbf{w}$ (usually a Gamma prior $\boldsymbol{\Gamma}(\mathbf{w})$), which favors most of the weights being set to zero.

Rather than or in addition to a sparseness prior, a new graphical prior be employed, *utilizing all available data* from $\mathbf{D}_\mathrm{L}$ and $\mathbf{D}_\mathrm{U}$ (labeled and unlabeled data, respectively). Figure 3 shows a graphical rendering for handwritten-digit recognition, as an illustrative example. The distance between any two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined via a general metric, where here we employ a kernel $K_\mathrm{o}(\mathbf{x}_i,\mathbf{x}_j)$. Note that this need not be the same kernel as employed in the classifier in (1). For any function $g(\mathbf{x})$, we introduce the cost function

$$C = \sum_{i=1}^{N_L+N_U} \sum_{j=1}^{N_L+N_U} K_\mathrm{o}(\mathbf{x}_i,\mathbf{x}_j)[\mathrm{g}(\mathbf{x}_i) - \mathrm{g}(\mathbf{x}_j)]^2 \qquad (2)$$

and we wish to minimize C, over the choice of functions $g(\mathbf{x})$. We sum over all available feature vectors: N_L labeled examples and N_U unlabeled examples. Note that when the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is small, the kernel $K_\mathrm{o}(\mathbf{x}_i,\mathbf{x}_j)$ is large, thereby forcing $\mathrm{g}(\mathbf{x}_i) \approx \mathrm{g}(\mathbf{x}_j)$. However, when $\mathbf{x}_i$ and $\mathbf{x}_j$ are dissimilar $K_\mathrm{o}(\mathbf{x}_i,\mathbf{x}_j)\rightarrow 0$, and therefore there is little restriction placed on $\mathrm{g}(\mathbf{x}_i)$ with respect to $\mathrm{g}(\mathbf{x}_j)$. This implies that targets/non-targets that are similar in feature space should be classified similarly. It can be shown that minimizing C in (2) is equivalent to choosing $\mathbf{g} = \{g(\mathbf{x}_1),\ g(\mathbf{x}_2)\ ,...,g(\mathbf{x}_{N_L+N_U})\}^\mathrm{T}$ that minimizes $\mathbf{g}^\mathrm{T}\boldsymbol{\Delta}\mathbf{g}$, where $\boldsymbol{\Delta}$ is the graph Laplacian, defined as $\boldsymbol{\Delta} = \mathbf{D} - \mathbf{K}_\mathrm{o}$. The ijth element of $\mathbf{K}_\mathrm{o}$ is $K_\mathrm{o}(\mathbf{x}_i,\mathbf{x}_j)$ and $\mathbf{D}$ is a diagonal matrix, the ith element of which is $D_{ii} = \sum_j K(\mathbf{x}_j,\mathbf{x}_i)$.

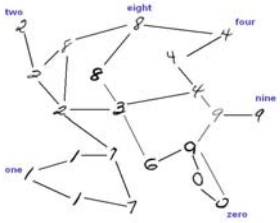

Figure 3. Illustration of a graph, in which the connections are dictated by similarity, defined by a kernel distance measure.

The above analysis is true for any function $g(\mathbf{x})$, and we now consider the special case $g(\mathbf{x})=f(\mathbf{x})$, where $f(\mathbf{x}) = \mathbf{w}^\mathrm{T}\boldsymbol{\varphi}(\mathbf{x})$ as defined in (1). In this case we have $\mathbf{g} = \boldsymbol{\Phi}\mathbf{w}$, where $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}(\mathbf{x}_1),\ \boldsymbol{\varphi}(\mathbf{x}_2),\ ...,\boldsymbol{\varphi}(\mathbf{x}_{N_L+N_U})\}^\mathrm{T}$. Therefore, for our problem we must choose the $\mathbf{w}$ that minimizes $\mathbf{w}^\mathrm{T}\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Delta}\boldsymbol{\Phi}\mathbf{w}$, which is equivalent to choosing the $\mathbf{w}$ that maximizes the zero-mean Gaussian random field prior with covariance matrix $\boldsymbol{\Sigma} = [\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Delta}\boldsymbol{\Phi}]^{-1}$, $\mathrm{p}(\mathbf{w}|\mathbf{D}_\mathrm{L} \cup \mathbf{D}_\mathrm{U}) \sim N(\mathbf{0},\boldsymbol{\Sigma})$.

Within the training process, for determination of the classifier weights $\mathbf{w}$, the overall algorithm reduces to a maximum *a posteriori* (MAP) estimate

$$\mathbf{w}_{\mathrm{MAP}} = \max_{\mathbf{w}} \left\{ \underbrace{\prod_{n=1}^{N_L} p(y_n \left| \mathbf{x}_n, \mathbf{w}) \right.}_{\text{Labeled Training Data, } \mathbf{D}_\mathrm{L}} + \underbrace{\lambda_\mathrm{G} \mathrm{p}(\mathbf{w} \left| \mathbf{D}_\mathrm{L} \cup \mathbf{D}_\mathrm{U}) \right.}_{\text{Graphical Prior Utilizing } \underline{\textit{All}} \text{ Available Data } \mathbf{D}_\mathrm{L} \& \mathbf{D}_\mathrm{U}} + \underbrace{\lambda_\mathrm{P} \Gamma(\mathbf{w})}_{\text{Sparseness Gamma Prior}} \right\} \qquad (3)$$

The relative importance of the three terms above is controlled by the Lagrange multipliers λ_G and λ_P, which are optimized in the training phase. The framework indicated above, first developed at Duke under 6.1 ONR support, has an attractive interpretation: The first term controls the impact of the labeled training data $\mathbf{D}_\mathrm{L}$; the second term accounts for *all* available data $\mathbf{D}_\mathrm{L}$ and $\mathbf{D}_\mathrm{U}$, via the aforementioned graphical prior; and the last term imposes sparseness via a Gamma prior. The sparseness selects only the most informative (relevant) feature vectors for classifier design. This same type of sparseness prior may also be used to select the most informative feature components.

III.1.2. Feature selection

The framework discussed above is naturally suited for investigation of feature selection. We may employ a linear kernel $f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^\mathrm{T}\mathbf{x} = \sum_{i=1}^{N_f} w_i x_i$, where x_i represents the ith feature component and N_f is the total number of features. The sparseness Gamma prior $\Gamma(\mathbf{w})$ favors a classifier for which most w_i are zero, and therefore this prior favors a solution for which most of the feature components x_i are ignored. In this manner, through training, we learn which of the features are most important for the classification task. The important thing to note is that, via the graphical prior $\mathrm{p}(\mathbf{w}|\mathbf{D}_\mathrm{L} \cup \mathbf{D}_\mathrm{U})$, the important features are selected using all available data, labeled and unlabeled. This is a critical characteristic in the context of adaptivity, as discussed further below. This framework will be extended such that the most relevant feature vectors and feature components are selected simultaneously, via appropriate sparseness priors.

Extracting robust features from the sensor information will be explored using Quadratic Correlation Filters (QCFs). Linear correlation filters have been used successfully for addressing the target detection problem. It is however often necessary to use a large number of linear filters for dealing with challenging detection problems. The disadvantage is that each linear filter is designed to operate independently of the others. As a result the correlation surface of all filters must be individually searched and a winner must be selected. A major benefit of QCFs may be to reduce the processing complexity by requiring fewer overall computations and simplifying the decision process. It can be shown that although the QCF requires several parallel linear correlations to implement, the linear branches work together to implement one QCF, and their outputs are combined into a single detection output. This greatly reduces post-processing complexity as the need to search many separate correlation surfaces is eliminated. Another advantage of quadratic filters is that they are able to better exploit the second order statistics of the data. It is well known that quadratic classifiers are optimum for

Gaussian distributions. However, even when the data is not necessarily Gaussian we expect QCFs to perform better than their linear counterparts, under general conditions.

The issue of finding targets in background clutter is a two-class pattern recognition problem. Consider an input signal $\mathbf{x} = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \end{bmatrix}^T$ that can be either a target (class ω_1) or background (class ω_2). For now we assume that $\mathbf{x}$ is a purely real signal. The output of the quadratic filter is defined as

$$y = \mathbf{x}^T \mathbf{T} \mathbf{x} = \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} x(i) x(j) \qquad (4)$$

The coefficient matrix $\mathbf{T}=\{t_{ij}\}$ is square and real but otherwise unrestricted. It should be noted that since $\mathbf{T}$ is not restricted to be positive definite, the output y can be either positive or negative. The idea is to determine $\mathbf{T}$ such that y is positive and as large as possible when $\mathbf{x} \in \omega_1$, and is negative or as small as possible when $\mathbf{x} \in \omega_2$.

LMM&FC has developed extensive methodologies for estimating the matrix $\mathbf{T}$, based upon separation of information and class separation. In either case, the problem reduces to projecting the data onto a subspace (of reduced dimension) where separation of classes is done optimally. The class separation metric leads to a subspace defined by the eigenvectors of the matrix

$$\left(R_X + R_Y\right)^{-1}\left(R_X - R_Y\right) \qquad (5)$$

where R_X and R_Y are the correlation matrices for the training data of class X and Y respectively. Since we have a subspace defined, we have basis vectors which determine this subspace. The projections of the data onto these basis vectors become features for classification. It has been shown by experimentation, that since the basis vectors are derived from optimizing a quadratic form, the optimal discriminate function is quadratic. Thus, the energy metric of (4) becomes optimal.

It is envisioned that coupling the QCF architecture with unlabeled data will lead to higher-order decision boundaries. Thus, extending the QCF architecture to unlabeled data via the graph-based similarity metric depicted by Fig. 1 holds promise in extracting robust features for further classification algorithms, outlined in Sec. V.1.1.

III.1.3. Adaptivity and sensor management

The Bayesian formalism discussed in Sec. V.1.1 allows several strategies by which the classifier may adapt to its environment and to the characteristics of targets. Note from (3) that the classifier utilizes the labeled training data $\mathbf{D}_L$ and also all of the available unlabeled data $\mathbf{D}_U$ (which is to be labeled or classified). In this manner, as the environmental and target characteristics change (*i.e.*, as $\mathbf{D}_U$ changes) the *algorithm automatically refines itself* via the prior $p(\mathbf{w}|\mathbf{D}_L \cup \mathbf{D}_U)$. For example, the algorithm may refine what are deemed to be the most appropriate features, as environmental conditions change.

We also utilize new techniques in the design of optimal experiments (DOE), whereby we *integrate sensing and ATR*. Specifically, rather than simply applying the algorithms to whatever data the sensor collects, the algorithm will guide selection of new multi-sensor data (defining new data for $\mathbf{D}_{\mathrm{U}}$). As a result of this process the parameters of the classifier are refined, as are detection thresholds, based on (3).

Providing more details on the DOE technique, recall that the data-dependent component of the Hessian, as defined on the posterior likelihood of the weights **w**, is expressed using (3) as

$$H_{ij}(\mathbf{D}_{\mathrm{L}} \cup \mathbf{D}_{\mathrm{U}}) = \frac{\partial}{\partial w_i}\frac{\partial}{\partial w_j}\left\{\sum_{n=1}^{N_L} p(y_n|\mathbf{x}_n, \mathbf{w}) + \lambda_{\mathrm{G}} \mathrm{p}(\mathbf{w}|\mathbf{D}_{\mathrm{L}} \cup \mathbf{D}_{\mathrm{U}})\right\}$$

By taking the determinant of the Hessian matrix, we quantify the information content in the data $\mathbf{D}_{\mathrm{L}}$ and $\mathbf{D}_{\mathrm{U}}$ (within a second-order – Gaussian – approximation), in the context of learning the classifier weights **w**. Considering now the labeled and unlabeled data, $\mathbf{D}_{\mathrm{L}} = \{\mathbf{x}_n^{(m)}, y_n : \forall\ m \in S_n\}_{n=1,N_L}$ and $\mathbf{D}_{\mathrm{U}} = \{\mathbf{x}_n^{(m)} : \forall\ m \in S_n\}_{n=N_L+1,N_L+N_U}$, we may ask the following question: Given the available unlabeled data $\mathbf{D}_{\mathrm{U}}$, which sensor measurements should be performed that would best improve our ability to classify? For example, if we have feature vectors $\mathbf{x}_n^{(m)}$ for sensors m in the set S_n, we may ask which sensors $m \notin S_n$ should be deployed to acquire new features (physics) for a given target.

For sensing paradigms for which a given target may be viewed from multiple orientations, the hidden Markov model (HMM) is a natural tool for processing the data and for performing optimal design of experiments (see Subtask 3.2). In Sec. V.1.5 we provide further details on the HMM and on how the HMM may be placed within a semi-supervised setting.

III.1.4. Integrating multiple information sources: Bayesian co-training

The techniques summarized above address semi-supervised utilization of labeled and unlabeled data, via graph-theoretic techniques. We now discuss how this graph-based statistical prior may be utilized to integrate multiple information sources. The framework in (2) dictates that if a given source is characterized by feature vector **x**, and if $\mathbf{x}_i$ and $\mathbf{x}_j$ are closely connected within the graph, feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ should yield similar classification decisions. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are not connected graphically, it is deemed statistically unlikely that they are associated with the same target/clutter class. We now consider an additional (separate) information source, characterized by the feature vector **z**. In this case we have three scenarios: (i) items characterized only by information source **x**, (ii) items characterized only by information source **z**, and (iii) items characterized by both information sources **x** and **z**. This addresses the likely situation for which the information associated with **x** may be available for some items, the information associated with **z** may be available for other items, and a subset of items may have features from both information types. We now seek to integrate this disparate information.

This problem will be addressed as follows. Two graphs will be designed, using (2), one for items with feature vector **x** and a second for items with feature vector **z**. We now seek to couple these graphs, using those items for which both information sources **x** and **z** are available. As in (1), a function $f_{\mathrm{x}}(\mathbf{x}|\mathbf{w}_{\mathrm{x}})$ is defined on the graph associated with feature-vector **x**, and a separate function $f_{\mathrm{z}}(\mathbf{z}|\mathbf{w}_{\mathrm{z}})$ is defined on the graph associated with **z**. For those items for which information sources **x** and **z** are both available, we enforce the condition that it is statistically likely that $f_{\mathrm{x}}(\mathbf{x}|\mathbf{w}_{\mathrm{x}})$ and $f_{\mathrm{z}}(\mathbf{z}|\mathbf{w}_{\mathrm{z}})$ yield similar results (*i.e.*, a statistical prior should be designed that favors that these two information sources yield a similar classification decision). Let the set $\mathbf{S}_{\mathrm{B}}$ represent those examples for which both information sources are available. As an extension of (2), we add the additional condition that the following should be minimized

$$C_{\mathrm{B}} = \sum_{i \in \mathbf{S}_{\mathrm{B}}} \sum_{j \in \mathbf{S}_{\mathrm{B}}} [f_{\mathrm{x}}(\mathbf{x}_i|\mathbf{w}_{\mathrm{x}}) - f_{\mathrm{z}}(\mathbf{z}_j|\mathbf{w}_{\mathrm{z}})]^2 \qquad (6)$$

This yields a joint prior on the weights $\mathbf{w}_{\mathrm{x}}$ and $\mathbf{w}_{\mathrm{z}}$, linked where both information types are available. Algorithmically, a new term is added to (3), accounting for (6). Similar to the terms in (3), the relative importance of the new term is controlled by a Lagrange multiplier λ_{B}, learned when training. In the machine-learning community utilization of multiple "views" of the same item (here multiple information sources) is termed "co-training". The above formalism yields Bayesian co-training, first developed at Duke under ONR 6.1 support. We will apply this technique to the asymmetric-warfare problems of interest, while also accounting for an arbitrary number of information sources (not only two information sources, **x** and **z**, discussed above for simplicity).

III.1.5. A collaborative ATR framework

Consider the scenario in which an object has been detected, and we wish to verify its class. To simplify the discussion, assume that two separate ATRs (that produce a 2-bit code) must be designed to work together to recognize two different classes denoted by ω_x and ω_y. For the purposes of discussion, we treat each ATR to be a MACH type correlation filter, and represent them as $H_1(k,l)$ and $H_2(k,l)$, respectively. We require that if class ω_x is present, $H_1(k,l)$ should produce a large positive output which is treated as a "1" if it exceeds a threshold T_1. Similarly, $H_2(k,l)$ should produce a large negative output which is treated as a "0" if it less than a threshold T_2. Thus, the output code [1 0] should be obtained whenever ω_x is present. Conversely, if ω_y is present, the filters are designed such that $H_1(k,l)$ yields a large negative value while $H_2(k,l)$ yields a large positive value, producing the output code [0 1]. Using these two ATR algorithms and their outputs, how may we optimize the sensors relative to the target?

We seek a metric that characterizes performance as a function of viewing geometry, and then drive the configuration of the sensors to optimize the performance metric. Towards this end, we define a *distance* or *separation* metric based on the MACH filter algorithm (a similar function may be derived for essentially any ATR algorithm). Let $X_i(k,l)$

represent the 2-D Fourier transforms of N training images of class ω_x, selected to represent viewing angles of the class 1 object around θ°. Similarly, $Y_i(k,l)$ are the 2-D Fourier transforms of N training images of class ω_y that represent viewing angles of the class 2 object around α°. The *mean* and *spectral variance* for each of the classes are defined as

$$M_x^\theta(k,l)=\frac{1}{N}\sum_{i=1}^{N}X_i(k,l) \qquad S_x^\theta(k,l)=\sum_{i=1}^{N}\left|X_i(k,l)-M_x^\theta(k,l)\right|^2$$

$$M_y^\alpha(k,l)=\frac{1}{N}\sum_{i=1}^{N}Y_i(k,l) \qquad S_y^\alpha(k,l)=\sum_{i=1}^{N}\left|Y_i(k,l)-M_y^\alpha(k,l)\right|^2$$

We first discuss the design of one of the filters, say $H_1(k,l)$; the design of the second filter will follow the same paradigm. The expression for the first MACH filter that separates a θ° view of class 1 from a α° view of class 2 is

$$H_1(k,l)=\frac{M_x^\theta(k,l)-M_y^\alpha(k,l)}{S_x^\theta(k,l)+S_y^\alpha(k,l)}$$

In fact, it is easy to show that *distance* or *separation* produced by this filter as function of the angles α and θ is given by

$$Q(\theta,\alpha)=\sum_k\sum_l\left|\left[M_x^\theta(k,l)-M_y^\alpha(k,l)\right]^* H_1(k,l)\right|^2=\sum_k\sum_l\frac{\left|M_x^\theta(k,l)-M_y^\alpha(k,l)\right|^2}{S_x^\theta(k,l)+S_y^\alpha(k,l)}$$

We refer to this function as the *MACH separation metric.* Our strategy is to train the filter at the specific viewing angles of each class that maximize $Q(\theta,\alpha)$.

We assume that ATR-1 and ATR-2 are on separate platforms, each with its own sensor. To drive the relative position of the sensors in an optimum manner consistent with a MACH filter class separation metric, the platforms must move in the specific formation which maximizes Q. Under the hypothesis that the object belongs to Class-1, ATR-1 should yield a strong positive response (code bit 1) when the object is viewed from angle θ_1. Similarly, ATR-2 should produce a strong negative response (code bit 0) when the object is viewed from the angle θ_2. If the orientation of the object is known, both ATR-1 and ATR-2 can move to the necessary positions and obtain images at the optimum angles. Otherwise, the two platforms should move around the object with a relative angular separation of $\theta_1-\theta_2$, checking to see if a strong [1 0] code is obtained. Similarly, to verify the hypothesis that the object belongs to Class-2, the sensors should group into a new formation with a relative angular separation of $\alpha_1-\alpha_2$ and move around the object to see if the code [0 1] is obtained.

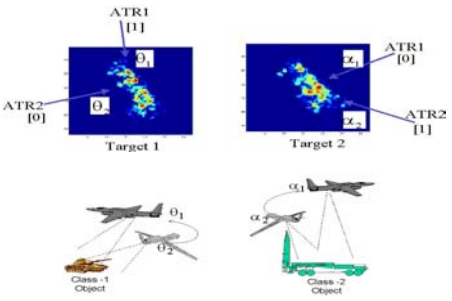

Figure 4. Adaptive optimization of the sensor position relative to a target, using collaborative ATR.

The collaborative ATR scenario is built upon local processing of each sensor, with the final outputs delivered to a central command console in the form of an ATR codeword. It is envisioned that this type of collaborating ATR can offer a final classification decision, or offer the formal labeling of the clusters which make up $\mathbf{D}_U$. The framework will be implemented in a semi-supervised setting (see Sec. V.1.1), thereby accounting for limited labeled (training) data and large quantities of unlabeled data $\mathbf{D}_U$.

III.1.6. Extension to video

Feature optimization, information integration and sensor management have been discussed in the context of state-of-the-art kernel-based algorithms. For time-dependent data, *e.g.*, video, it is desirable to extend these ideas to algorithms that explicitly exploit time-dependent phenomena. For such data we will therefore employ a hidden Markov model (HMM), with such widely employed for time-dependent data such as speech. The key new contribution to be pursued involves integrating the HMM within a semi-supervised framework. Recall that the semi-supervised construct is motivated by the fact that in the asymmetric-warfare problem of interest, we are unlikely to have a large quantity of labeled data $\mathbf{D}_L$, while there is likely to be much unlabeled data $\mathbf{D}_U$. For example, in urban warfare, it is unlikely that there will be a large quantity of data $\mathbf{D}_L$ consistent with suicide-bomber behavior. However, there is likely to be very large quantities of data consistent with general human behavior in an urban environment, this collected as a given environment is being monitored. The objective is to design an HMM classifier based on the small quantity of labeled data *and* the large quantity of unlabeled data. Moreover, the classifier should not characterize a given sequence in isolation, but rather should place that video in context, based on all other data available from a given environment. Traditional HMMs are purely supervised, in that they are trained exclusively on the labeled data $\mathbf{D}_L$.

In the semi-supervised HMM, the unlabeled data $\mathbf{D}_U$ is integrated with the labeled data $\mathbf{D}_L$ via a modification of the traditional expectation-maximization (EM) algorithm. In particular, in the classical *supervised* EM training of an HMM, the hidden variables are the states S in which a given feature resides (recall that the HMM is characterized by a underlying hidden state sequence, modeled as a Markov process). In the *semi-supervised* HMM, we introduce a new hidden variable: the label y associated with a given unlabeled sequence (the label characterizes the different types of conditions characteristic of a given video sequence). Duke has demonstrated, in its 6.1 ONR research, that this relatively modest extension of the EM algorithm (introduction of a new hidden variable), allowing semi-supervised HMM training, yields markedly improved classification performance.

In addition to developing the semi-supervised HMM to perform classification, we will utilize the HMM within the optimal design of experiments, to optimize the operation of video sensors. This will be performed as follows. Rather than making a Laplace approximation to the HMM parameters, as done in Sec. V.1.3 for the kernel machines, a variational approach will be employed to characterize the likelihood of HMM parameters given previously observed data. Once the posterior likelihood of HMM parameters is so computed, the selection of which video to acquire next proceeds analogous to the

procedure discussed in Sec. V.1.3. Specifically, we perform that measurement that most reduces the *expected* entropy in the HMM parameters.

Some of the challenges to be pursued include coupling video data with other information sources (still images, waveforms and human intelligence). This will involve an integration of kernel-based classifiers and HMMs within a single paradigm. Toward this end, the statistical output of the kernel classifier, defined in Sec. V.1.1, will be coupled with the natural statistical output of the HMM.

III.2 Results targeted

The research is targeted primarily toward asymmetric warfare, urban warfare, and port/base security. In consultation with ONR, and utilizing the significant experimental resources of NAVAIR and LMM&FC, measurements will be performed to provide a data base for algorithm development and refinement. The data will be targeted toward situations of: (i) significant quantities of unlabeled data $\mathbf{D}_U$; (ii) small and possibly no labeled data $\mathbf{D}_L$; and (iii) multiple sensor modalities, including video. We will also utilize existing data from previous extensive collections performed by NAVAIR and LMM&FC. It is our goal to demonstrate the following targeted objectives:

- Within an environment of limited labeled data and large quantities unlabeled data, from multiple information sources, extract the most informative data features, accounting for the classification task. We hope to demonstrate an autonomous pipeline from data to information, thereby significantly reducing burdens on the human analyst.
- Semi-supervised detection of low-likelihood events, for which limited labeled training data may be available (*e.g.*, suicide attackers at a port/base), exploiting multiple sensor modalities, including video.
- Optimization of all information resources, using multiple sensor modalities as well as HUMINT. This will be demonstrated using graph-based semi-supervised algorithms, within the context of optimal experiments and collaborative ATR.

IV. Operational Utility

This research program is a coordinated effort between NAVAIR, SIG and LMM&FC. In this section we discuss the operational utility of the entire program.

A high-level concept of how the video and signal management module would integrate in the FORCEnet architecture is illustrated in Fig. 5. To integrate sensors, networks and decision aids with a client, all combatants within the theater of operation (as well as many non-combatants outside it) must be able to communicate with each other, receiving and transmitting updated information over a network backbone. The architecture as presently envisioned assumes an agent-based computing infrastructure. In fact FORCEnet is built on the concept of an agent-based computing infrastructure that will provide a dynamic, highly reconfigurable command-and-control system to contain large numbers of heterogeneous sensors and systems. The CoABS research community, a DARPA initiative, is developing a prototype agent grid as an infrastructure for the run-time integration of heterogeneous multi-agent and legacy systems. Since as advertised "The CoABS Grid is middleware that integrates heterogeneous agent-based systems, object-based applications, and legacy systems", they provide a software vehicle to transport algorithms to sensors (*e.g.,* Java Script) and a foundation to wrap legacy systems. The notional model in Fig. 5 also assumes that some language like Extensible Markup Language (XML) or even the DARPA Agent Markup Language (DAML) that is being developed as an extension to XML to process the metadata tags will be available on the network. Thus, an agent's output or service is implicitly available to (i) initiate a request for service from the archive or (ii) broker a catalog algorithm delivery to a client, but its final instantiation is not critical to this research.

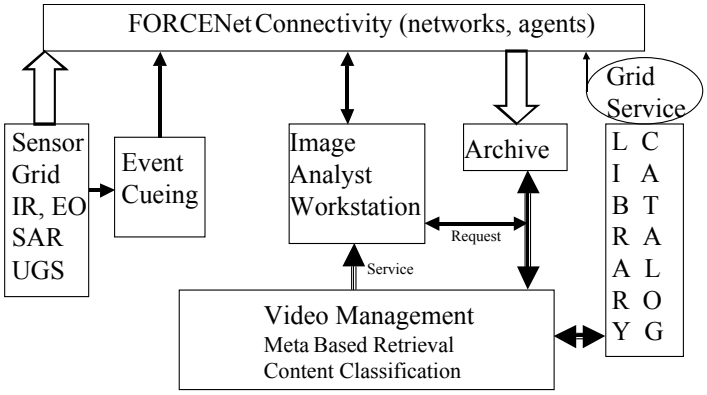

Figure 5. Video signal management notional interconnection.

The signals or images will reside in a tiered architecture that will provide distributed surveillance in the air, sea and land domains. The sensor grid components will include unattended ground sensors (UGS), as well as sensors residing on lower-tier manned and unmanned aircraft, sea-borne vessels and upper-tier assets. UGS are signal sources, while synthetic aperture radar (SAR), laser radars, sensor arrays, electro-optical and infrared cameras are image sources. An organic or distributed event monitoring capability based on local unusual sensor patterns, possibly using an agent service will initiate cueing messages.

In terms of describing the operational utility of this work, let us consider operational situations ranging from the simplest and least stressing to the most challenging, in

keeping with the crawl, walk, run and fly paradigm. The initial work will be geared to address the problem of the image analyst operating in the Marine Tactical Exploitation Group (TEG) or perhaps a sailor aboard a carrier utilizing the Tactical Exploitation System as part of DCGS, the Distributed Common Ground Station. The analyst is currently presented with a single stream of imagery to assess and determine possible target locations, and will be asked in the future to ingest and process multiple data streams from distributed heterogeneous sensors. Under this program we will automate the target-detection process within ATR algorithms, increasing the number of potential targets that can be passed through the targeting chain, reduce the operational workload of the image analyst, and increase available time for analysts to consider difficult imagery where the target detection algorithms will not provide the required performance. In this scenario, the advantage is clear, providing a suitable metric in terms of the number of detected targets, both by automatic systems and the analyst. From the assessment studies, the false-alarm rates of the system will be documented, but in the operational sense a "frustration" metric will need to be developed. If the false-alarm rate is too high, at some point analysts will refuse to use the system, as it generates too many false alarms resulting in increased analyst workload rather than lowering it, while if the false-alarm rate is too low, we risk lowering the probability of detection to the point that human analyst will consistently find more targets. The feedback from an image analyst can be used to drive development of labeled and unlabeled data sets to adapt the classifiers.

Progressing to "walk", the decision-making process moves forward from a central location to an airborne asset, such as the Hairy Buffalo, an experimental P-3 testbed. The P-3 houses an APY-6 radar system, imaging infra-red sensors, a multi-spectral push-broom sensor, and plans exist to incorporate a Tactical Control System (TCS) for controlling a UAV. In this case, the targeteer aboard the Hairy Buffalo will need to be able to ingest multiple data streams, and detect targets in a computational and bandwidth constrained environment. The same metrics used for a centrally located ground based image analyst apply in this scenario. By moving the decision-making process further forward, reduced timelines from target detection to target prosecution will result.

Having developed and assessed new multidimensional target detection approaches for networked sensors, the approach enables the progression to "run". Under the FORCEnet concept, the battlefield will be populated by large numbers of heterogeneous sensors, both imaging and non-imaging, which will provide persistent surveillance of the battlespace. Unless technology advances to enable some measure of autonomy and cooperation among the sensor grid, image analysts and system operators will be overwhelmed by a virtual avalanche of digital data. Just one enhanced resolution color video system proposed by National Imagery and Mapping Agency (NIMA) operating with 1280×720 pixels (progressive scan) at 60 frames/sec and 10-12 bit dynamic range will generate a raw stream of about 0.8 terabytes / hour. The theory-of-optimal-experiments approach will provide some measure of autonomy and cooperation, by attempting to make the best possible use of the global information available on the ESG – determining which data would be best to consider next to provide an optimal decision. Data availability will most likely be determined by use of an agent-based computing architecture, such as the one explored in the DARPA CoAbS program. Given the data

availability, the information-theoretic (design of experiments) approach will request data and possibly task other sensors to further process potential target detections. Only when a specified data quality assurance level has been achieved will an Event Cue as seen in Fig. 5 be generated, reducing image analyst workload, reducing bandwidth demands, and maximizing information usage, providing the highest targeting throughput with minimal false alarms.

V. Year 1 Results

IV.1. Video Analytics

The Program Manager, Dr. Shubha Kadambe, visited the SIG offices on August 11, 2006, during which SIG presented a detailed set of results on several video data sets associated with complex scenes. The presentations and corresponding video were given to Dr. Kadambe on CD; we here concentrate on summarizing the technical approaches, since the video itself clearly cannot be shown.

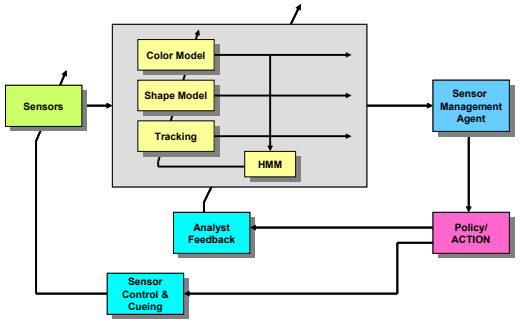

Figure 6. Overall summary of video processing scheme.

In Fig. 6 we summarize the overall video-processing algorithm. Adaptive tracking of general moving objects is performed by developing a Gaussian mixture model (GMM) for each pixel in the scene. The GMM is continually updated, to adaptively learn the properties of the environment. When a pixel is deemed to be of low likelihood by the background model, it is characterized as foreground. Proximate foreground pixels are then aggregated (via simple foreground-pixel adjacency analysis). The motion of the foreground object is estimated, and in a Bayesian sense we predict the next shape and position information, this serving as a prior for the next frame. This component plays a critical role in tracking moving objects through occlusions, with this demonstrated on several complex video data sets. A hidden Markov model (HMM), which learns typical behavior, is used to help track the foreground objects. All of these components have been completed successfully in Year 1, for arbitrary video sequences.

In Year 2 we will focus on the sensor-management agent, which will utilize reinforcement learning (RL) technology to learn a policy for the video camera(s). The RL algorithm will develop a real-time policy for control of the sensors. For example, if something anomalous is apparent, it may be desirable to zoom the camera, to obtain finer-resolution data. However, this is done at the cost of losing the ability to see the broader scene. Therefore, it may be desirable to augment the parameters of other available sensors (e.g., other cameras) to "cover" for the camera performing zooming. The cost-sensitive sensor-management agent develops a policy that maps sensor observations to optimal sensing actions, accounting for the cost of the action and the associated risk to the decision maker (Bayes risk).

As indicated in Fig. 6, one possible action may be to employ an analyst, and ask for a label for a given scene (with binary label benign or dangerous). There are two salutary aspects of such use of analysts. First, by focusing the analyst only on those aspects of the scene for which labels would be most informative to the algorithm, the analyst workload is reduced. Hence, by optimally integrating the analyst with the algorithm, overall system performance is improved while collectively allowing all video to be analyzed (by the algorithm and analyst). In addition, use of the analyst also allows the algorithm to continually learn, and therefore refine itself to new scenes or to changes in a given scene. Let I^j represent the image frame at time *j*, A^j the labels for all pixels at time *j, t* the current time step (frame) and assume *n* frames of labeled training data. Then the inference with regard to the image to be viewed at frame *t*, given all previous frames $\vec{I}^{t-1}$ and training data $\vec{A}^n$ may be expressed rigorously as

$$P(I^t \mid \vec{I}^{t-1}, \vec{A}^n) = \sum_{\vec{A}^{t-1}} P(I^t \mid \vec{I}^{t-1}, \vec{A}^{t-1}) P(\vec{A}^{t-1} \mid \vec{I}^{t-1}, \vec{A}^n)$$

We now utilize the color (*C*), shape (*S*) and trajectory (*T*) information to constitute the representation

$$\begin{aligned} & P(I^t \mid \vec{I}^{t-1}, \vec{A}^n) \\ &= \sum_{\vec{A}^{t-1}} P(I^t \mid \vec{I}^{t-1}, \vec{A}^{t-1}) P(\vec{A}^{t-1} \mid \vec{I}^{t-1}, \vec{A}^n) \\ &= \sum_{\vec{A}^{t-1}} P(I^t \mid C^{t-1}, S^{t-1}, T^{t-1}) P(\vec{A}^{t-1} \mid \vec{I}^{t-1}, \vec{A}^n) \end{aligned}$$

Careful analysis of these equations reveals we need to compute and update three statistical quantities: $P(I^t \mid A^t, C^{t-1})$, $P(A^t \mid S^{t-1}, \mu)$ and $P(\mu \mid T^{t-1})$. The first term $P(I^t \mid A^t, C^{t-1})$ is a statistical mapping from current assignments (foreground/background) and previous pixel colors to future colors, with this modeled via a GMM. The term $P(A^t \mid S^{t-1}, \mu)$ is a mapping from the previous shape information and associated position to new labels. Finally, $P(\mu \mid T^{t-1})$ is essentially a tracker that estimates the location of the centroid of moving entities. This latter component is implemented via a simple Kalman filter, with future implementations utilizing a hidden Markov model (HMM). We present examples of each of these components, and their respective roles.

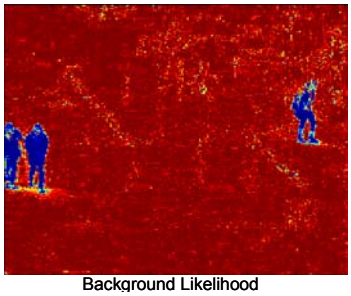

Figure 7. GMM extraction of foreground items from complex scene.

Considering first $P(I^t \mid A^t, C^{t-1})$, implemented via the GMM, we present an example result in Fig. 7. In this figure the red locates regions that are likely to be

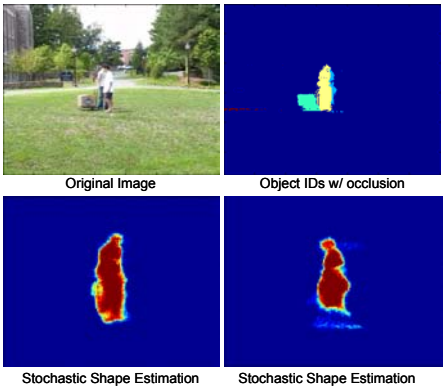

Figure 8. Frame from video (top left), and estimated shape of two moving people (bottom two figures). At top-right is shown the estimated identities (different colors) of moving entities. Note that the shape of the occluded person is maintained, through inference.

representative of background, given previous labels and colors. In this figure we note that the multiple moving objects (people) are accurately extracted from a complex scene.

Concerning $P(A^t|S^{t-1},\mu)$, a stochastic model is maintained for the label of each pixel, with this tied to previous shape information and predicted new location μ. The key aspect of this framework is that it maintains through inference the shape of multiple moving entities, even when occlusions occur, with a frame from a representative example presented in Fig. 8.

The final entity that must be statistically updated is $P(\mu|T^{t-1})$, which is estimated using a Kalman filter to keep track of the centroid of moving entities. An example of estimated tracks in video is shown in Fig. 9.

The results in Figs. 7-9, in addition to the large set of video examples given to Dr. Kadame, demonstrate that SIG has developed the tools to extract general moving entities from complex scenes. The algorithm extracts the moving entity, maintains an estimate of its shape, and can address multiple objects and associated occlusions. The next step, to be examined in detail in Year 2, is to develop statistical models of typical behavior (e.g., with an HMM), and to use such to detect anomalous behavior, of interest for asymmetric threats. As indicated in Fig. 6, the algorithm will be integrated within a sensor-management framework, with sensing policy learned adaptively via reinforcement learning. Note from Fig. 6 that a possible action of the algorithm is to ask an analyst for a label, thereby integrating the human into the adaptive algorithm learning, focusing the analyst on that data/video of greatest information for adaptive algorithm learning.

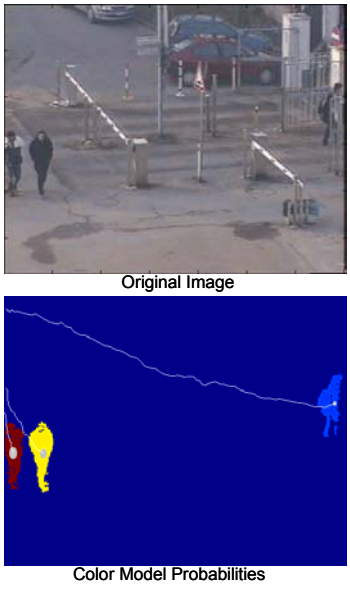

Figure 9. Tracked locations of three moving objects.

In Year 1 we have implemented preliminary algorithms to perform anomaly detection. Specifically, we have the ability to track the location of each item over time, through occlusions if necessary. We can therefore place a time constraint on the time a given item may remain in a given location; if this time period is exceeded, than a loitering event is detected. In addition, by maintaining a history of the tracks observed over a prescribed period (see Fig. 9), we can constitute a density function for the likelihood that a moving entity should be observed subsequently in a given area. This

Figure 10. Example video in which a loitering person is detected (blue) and an individual is located in a region not previously visited in data seen prior.

simple approach may be used to detect a moving entity in a location it shouldn't be. These concepts have been implemented in an adaptive, automatic algorithm, poised for immediate testing. In Fig. 10 we show an example of this algorithm, where here one individual has loitered too long and the other individual is walking in an area in which nobody has entered previously. Note that the algorithm maintains a count of the number of moving entities observed at all times, as well as the number of entities that are inferred to be occluded.

The algorithm has been carefully tested on a large set of video data, for variable numbers of types of moving entities, and for different background complexity. The performance appears to be quite promising. We note some of the algorithm strengths: (i) it employs a principled Bayesian framework, which is mathematically sound, with explicit assumptions and approximations, and it is capable of multiple hypotheses; (ii) the algorithm has proven to be robust, it being dynamic to changing environments, employing autonomous parameter calibration, and demonstrating a graceful failure mode and recovery; and (iii) it is powerful, since it incorporates strong color, spatial, and temporal dependencies, and it includes spatial and limited temporal information. The background removal relies on color information, naturally handling complex situations, e.g. occlusions, ambiguities, and it provides seamless expansion of additional evidence and models.

IV.2. Polynomial Correlation Filters and Multi-Frame Processing

Conventional correlation filters operate on a single image frame. PCFs are an extension where multiple image sources can be simultaneously and jointly filtered to produce a single correlation surface which is optimized with respect to all of the sources of data. We first provide an overview of the design of the multi-channel filter, and then illustrate its application and performance using multiple video image frames.

The notation followed in this section is as follows: images in the space domain are denoted in lower case italics while upper case italics are used to represent the same in the frequency domain. Thus, a two dimensional (2D) image $x(m,n)$ has Fourier transform $X(k,l)$. Vectors are represented by lower case bold characters while matrices are denoted by upper case bold characters. Either $x(m,n)$ or $X(k,l)$ can be expressed as a column vector $\mathbf{x}$ by lexicographical scanning. The superscript ^{T} denoted the transpose operation, and $^{+}$ denotes the complex conjugate transpose of vectors and matrices.

The output of the PCF can be mathematically expressed as

$$g_x(m,n) = \sum_{i=1}^{N} h_i(m,n) \otimes x^i(m,n)$$

where $x^i(m,n)$ is the image frame associated with the *i*-th channel, and $h_i(m,n)$ is the corresponding filter. The *N* filters jointly optimize a performance metric associated with the final output $g_x(m,n)$. An example of the PCF structure is shown in Figure 10 where the different channels represent non-linear functions of data from a given source although in general no restrictions are placed on what the channels or the sources of the inputs to the filters may be.

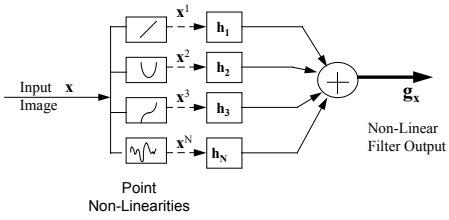

Figure 10. *N*-th Order Polynomial Correlation Filter Structure.

A sequence of video images of a van (the target) circling in a parking was collected using a COTS video webcam. A 64 x 128 region of each frame containing the target was extracted for training and testing purposes. Representative images from this set are shown in Figure 11. There were 512 frames collected covering 180 degrees of the target ranging from the front end view to the rear end-view (shown in the figure). The frames were separated by approximately 0.35 degrees.

Every other frame was chose as a training image for Channel 1. The input data for Channel 2 was arbitrarily selected to be the absolute difference between the channel 1 image and the view that lagged it by 50 frames (approximately 17.5 degrees separation). The 2-channel PCF was synthesized using these training images and then tested with the rest of the (non-training) images.

For testing purposes as well, the data input for Channel 2 was the absolute difference between the Channel 1 input image and a view that lagged it by 50 frames. For the purposes of comparison, the Channel 1 image was also processed by a conventional "single frame" *maximum average correlation height* (MACH) [1] filter synthesized using only the Channel 1 training images. The peak values were measured for both the 2-channel PCF and the conventional MACH filter and are plotted in Figure 12, and the cross-sections of representative correlation surfaces are shown in Figure 13.

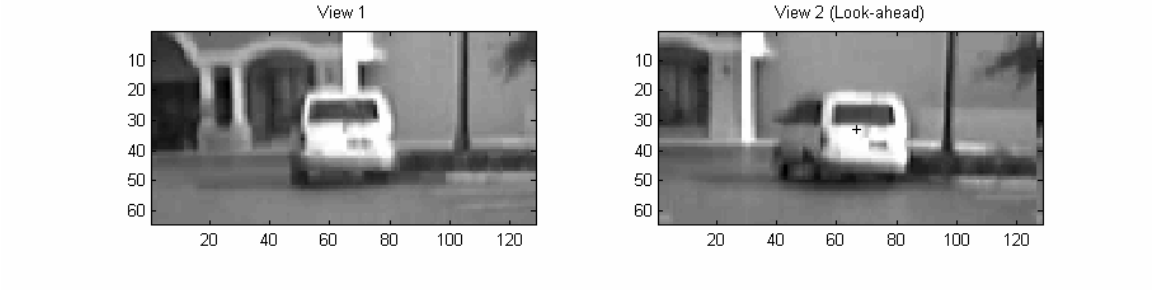

Figure 11. Representative images of the target at different points in the video sequence.

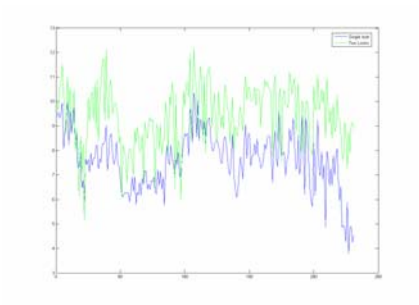

Figure 12. Plot of the peak correlation output for the 2-channel PCF (green) and the conventional single frame MACH filter (blue)

In Fig. 12 the vertical axis represents the peak value measured in terms of *peak to side-lobe ratio* (PSR) and the horizontal axis is the test frame number. It is clear that the 2-Channel PCF performs better than the single channel MACH filter at almost every frame. In fact, there is considerable performance improvements at most places with the exception of a few instances where the 2-look performance overlaps the single look result. We believe in these cases one of the channels had a poor contrast or illumination, resulting in the absence of extra information and thus the result was essentially the same as that for the single look case. Overall, the 2-look PCF and the single channel MACH filter outputs are separated by a Fisher ratio of over 3.5.

The cross-sections shown in Fig. 13 also show the advantages of two-look correlation filtering. Subjectively, the single look correlation result on the left is more noisy with larger side-lobes (false peaks) than the 2-look correlation result on the right. This accounts for the better PSR values and overall greater confidence in the correlation peak.

We now describe the approach where multiple correlation surfaces used in a Bayesian decision process to estimate target probability using a model for the target's motion. Correlation filtering is a popular approach for image-based automatic target recognition (ATR) due to several attractive properties. Four such properties are: (1) correlation filters are inherently shift-invariant, which makes them useful for applications in which target position is unknown; (2) they can be designed to tolerate variability in the target signature, which could be caused by such phenomena as in-plane and out-of-plane rotation, scale changes, thermal state variations, and configuration changes; (3) they have closed-form solutions, making the design stage efficient and predictable; and (4) they can be applied efficiently in the frequency domain using the fast Fourier transform (FFT) algorithm. Many advanced correlation filer designs are available which optimize certain criteria and/or focus on particular types of distortion. Also, while most designs assume linear filtering, non-linear designs such as quadratic and polynomial correlation filters are available. Thus, the field of correlation filters offers a wide variety of options to the user, who can select the filter design best suited to the particular application at hand.

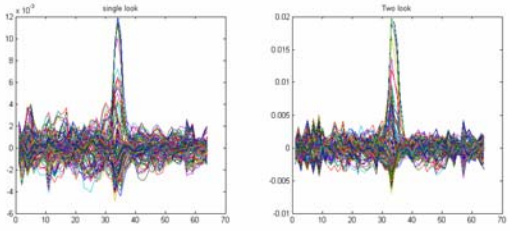

Figure 13. Comparison of the cross-section of the correlation surfaces obtained using (left) single-frame MACH filter and (right) 2-channel PCF.

In a typical ATR application, each value in the output of a correlation filter is compared to a pre-determined threshold, and the locations of all such values exceeding the threshold, called "peaks", may either be declared as target detections or fed into some sort of postprocessor (e.g., a tracker). When the images are corrupted by high amounts of noise, false peaks can occur frequently and may either cause too high a false alarm rate or confuse the tracking algorithm (if one is used). As an alternative to immediate thresholding in video-based ATR, we have developed a multiframe correlation filtering algorithm whereby correlation outputs are mapped to probability values and soft-information processing is performed in the mapped output *before* thresholding. A simple target motion model is employed by the algorithm to impose limitations on how targets may move in the scene; thus, noise-induced peaks which may have exceeded the threshold (and consequently mis-detected as targets in single-frame schemes) will not be classified as targets in the multiframe algorithm if they violate the assumed motion model.

Several advantages are realized by using this probabilistic approach. First, the algorithm avoids the information loss incurred by early thresholding, utilizing instead all of the available information in the outputs in a probabilistic sense. Second, the theory provides a means by which different correlation filters may communicate with one another, so that the likelihood of two conflicting detections at the same location is reduced. Third, we can weave into the algorithm probabilistic models of the terrain to disallow targets from spontaneously appearing in certain places (e.g., the middle of an empty field), which may help to further mitigate false detections. Most importantly, because the algorithm is not a substitute for existing correlation filter methods but rather an optional "attachment", we are able to retain all of the above-mentioned advantages of correlation filters in the multiframe algorithm, while potentially improving the recognition performance..

Our original multiframe correlation filter algorithm has several limitations including: (1) the ability to handle multiple targets was hindered by an assumed normalization constant which stood in for theory that was yet undeveloped; (2) we had not yet developed the theory that would allow multiple correlation filters to communicate with each other; and (3) there was no provision for a spatially varying target occlusion model. These problems have since been fixed by refining the underlying theory, resulting in a more robust algorithm.

In this report, we summarize the current status of the multiframe algorithm, describing in detail several improvements that have been made to the algorithm since our visit to Lockheed Martin. We point out some observations made during the visit that led to many of these improvements. We discuss not only the current capabilities of the algorithm but also potential applications and possible extensions.

Several limitations of the early version of the multiframe algorithm resulted from the method by which correlation values were mapped to probability values and subsequently combined with information from past frames. The correlation output from each frame was first mapped to an array of probability values using a predetermined mapping. Using the assumed motion model, a "prior probability" array corresponding to the next frame

was computed from the current array of probability values via a convolution. In the next frame, the same mapping was used to generate another probability array, and this array was combined with the computed prior probability array via a pointwise multiplication to form a "posterior probability array", or "enhanced array". During the creation of these probability arrays, it was assumed that the result of this pointwise multiplication should be normalized by a constant to yield true probability values. Because of the high complexity of the expression for this normalization constant, we approximated it instead of computing it directly.

As a consequence of the repeated multiplications and spatially invariant normalizations in the above algorithm, targets producing slightly stronger correlation responses tend to dominate and suppress those producing weaker responses after a long sequence of frames. This phenomenon was noticed in several test sequences containing multiple targets. Because of the resulting poor performance on these sequences, the normalization theory was revisited, and we discovered that the normalization actually should not be constant with respect to either position or time. *We derived new theory which includes an exact expression for the true normalization function*.

Upon examining this new theory, we further discovered that it was unnecessary to generate the intermediate probability array via the old mapping and subsequently combine it with the prior probability array. Instead, the posterior probability array could be computed by feeding both the correlation output and the prior probability array into a *single new* mapping function. The correct normalization is implicitly included in this new mapping function. *We have found that using the new mapping with implicit normalization results in much better handling of multiple targets*, i.e., a strong response from one target does not adversely affect the response of other targets. Preliminary results demonstrates the ability of the algorithm to handle multiple targets.

Previously the multiframe algorithm only supported a single correlation filter. Thus, if multiple filters were to be used in conjunction with the multiframe algorithm, each filter would need to be a separate "thread", i.e., the multiframe algorithm would be applied to the video multiple times in parallel, once for each filter. Such a scheme does not take advantage of the potential for communication between filters, where information from the output of one filter could be used in the calculating probabilities in the output of another filter. For example, if filters A and B were each designed to look for different target classes, we ought to be able to impose the constraint that the two filters cannot simultaneously detect a target at the same location, and the probabilities computed from their outputs should reflect this constraint.

The new mapping function described in the previous section is derived such that it simultaneously considers the outputs from as many filters as desired when computing each probability value. It also considers the entire group of prior probability arrays generated by the filters in the previous frame rather than just one such array. Communication between filters is thus implicitly carried out by this mapping function. With the advancements in the mapping function described above, information is allowed to flow more extensively than before, as illustrated schematically in Fig. 14.

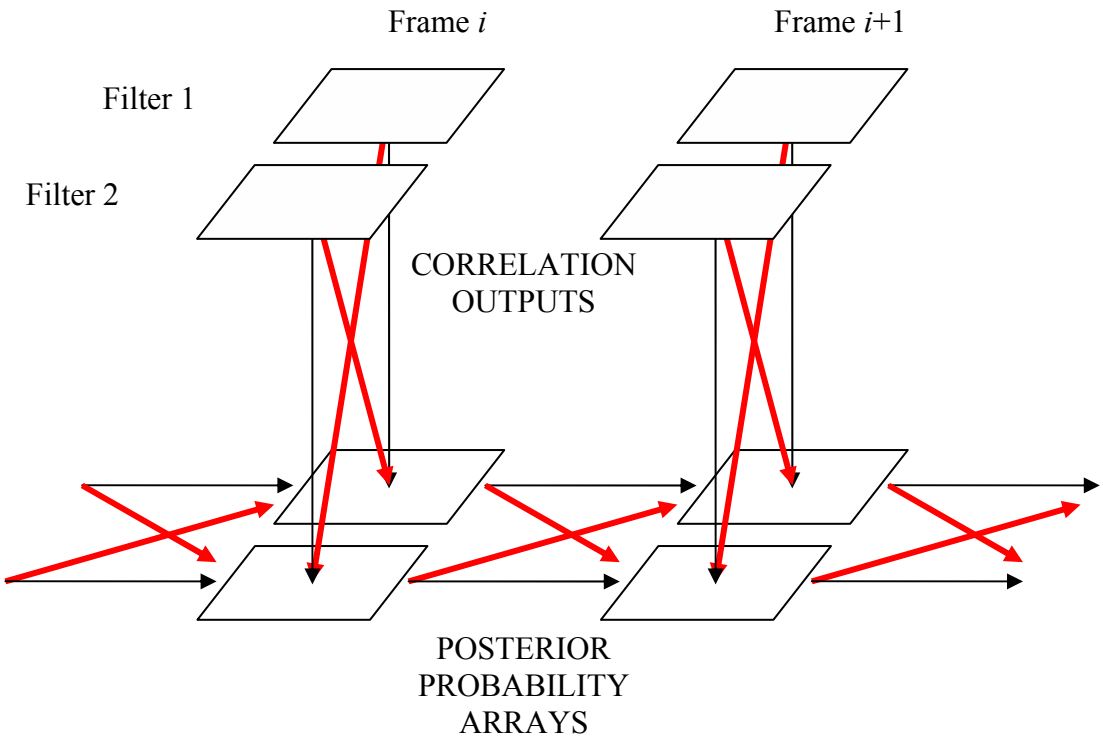

Figure 14. Information flow between correlation outputs and probability arrays in the multiframe algorithm (two-filter example). Thin black lines indicate flow in original algorithm, while thick red lines indicate additional flow achieved by improved mapping. Previously, there was no flow between separate threads (e.g., Filter 1 and Filter 2). After improvements to the theory, information from all previous prior probability arrays as well as all current correlation outputs is used to compute the posterior probability arrays. The intermediate prior probability arrays are not shown in this diagram.

VI. Transition Opportunities and Leveraging

The teaming of SIG with Lockheed Martin provides several teaming opportunities. Specifically, there are two key programs into which the ONR research may be transitioned:

Future Combat Systems (FCS)

Lockheed Martin is the prime contractor for Aided Target Recognition (AiTR) for FCS ground systems, for which the objective is to perform wide-area surveillance using FLIR and video sensors. The technologies developed under the ONR C2&CS program are directly relevant to FCS AiTR functions. The methods being developed here can be applied to detect and track moving vehicles and dismounts in surveillance imagery. Specifically, the algorithms may be included for evaluation in FCS Technology maturation activity.

DARPA's LACOSTE program

Lockheed Martin is one of the contractors selected to develop the LACOSTE sensor for persistent surveillance applications. The objective is to detect and track moving objects in FLIR images over a "ultra" large area. Phase II of LACOSTE program will require algorithms such as those developed under the ONR C2&CS program to exploit LACOSTE imagery, and to automatically adapt sensor parameters.

SIG will is also leveraging to programs:

AFRL Phase II SBIR

SIG has a Fast-Track Phase II SBIR directed toward multi-sensor base security. Integrian, a leading video-surveillance company, is an "investor" on this project, with this representing a drect transition opportunity to Integrian systems. Integrian is the contractor for video surveillance systems in New York City, Dallas, New Jersey, London, and Madrid, among many others. This also represents a significant opportunity for homeland security systems.

Video-based IED detection

SIG is applying similar techniques to those investigated here for the problem of video systems mounted on military vehicles, with this of significant importance for Marines (e.g., IEDs). SIG has just been issued a contract by NVESD to address this important problem.