Information-Theoretic Performance Analysis of Sensor Networks via Markov Modeling of Time Series Data

Yue Li, Devesh K. Jha, *Member, IEEE*, Asok Ray, *Fellow, IEEE*, and Thomas A. Wettergren, *Senior Member, IEEE*

***Abstract*—This paper presents information-theoretic performance analysis of passive sensor networks for detection of moving targets. The proposed method falls largely under the category of data-level information fusion in sensor networks. To this end, a measure of information contribution for sensors is formulated in a symbolic dynamics framework. The network information state is approximately represented as the largest principal component of the time series collected across the network. To quantify each sensor's contribution for generation of the information content, Markov machine models as well as x-Markov (pronounced as *cross*-Markov) machine models, conditioned on the network information state, are constructed; the difference between the conditional entropies of these machines is then treated as an approximate measure of information contribution by the respective sensors. The x-Markov models represent the conditional temporal statistics given the network information state. The proposed method has been validated on experimental data collected from a local area network of passive sensors for target detection, where the statistical characteristics of environmental disturbances are similar to those of the target signal in the sense of time scale and texture. A distinctive feature of the proposed algorithm is that the network decisions are independent of the behavior and identity of the individual sensors, which is desirable from computational perspectives. Results are presented to demonstrate the proposed method's efficacy to correctly identify the presence of a target with very low false-alarm rates. The performance of the underlying algorithm is compared with that of a recent data-driven, feature-level information fusion algorithm. It is shown that the proposed algorithm outperforms the other algorithm.**

***Index Terms*—Information fusion, sensor networks, symbolic time series analysis, target detection.**

Manuscript received February 22, 2017; accepted June 12, 2017. Date of publication July 6, 2017; date of current version May 15, 2018. This work was supported in part by the U.S. Office of Naval Research under Grant N00014-14-1-0545, and in part by the U.S. Air Force Office of Scientific Research under Grant FA9550-15-1-0400. This paper was recommended by Associate Editor Q. Ji. *(Corresponding author: Asok Ray.)*

Y. Li was with Pennsylvania State University, University Park, PA 16802 USA. He is now with FARO Technologies Inc., Orlando, FL 32746 USA (e-mail: yuesolo@gmail.com).

D. K. Jha was with Pennsylvania State University, University Park, PA 16802 USA. He is now with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 USA (e-mail: devesh.dkj@gmail.com).

A. Ray is with Pennsylvania State University, University Park, PA 16802 USA (e-mail: axr2@psu.edu).

T. A. Wettergren is with Pennsylvania State University, University Park, PA 16802 USA, and also with Naval Undersea Warfare Center, Newport, RI 02841 USA (e-mail: t.a.wettergren@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2717974

I. Introduction

SENSOR networks serve the important role of collecting and processing information in a variety of (possibly dynamic) environments, whose structure may be partially or completely unknown [1]. Such a network consists of a number of sensor nodes, each having the capability to process the signals that it measures from its own or a neighboring environment. The sensor nodes communicate with each other for information aggregation to enhance the quality of collected information in the network. However, due to the constraints on power, computation, and communication resources [2], [3], it becomes a challenge to reduce both installation and operating costs, while enhancing the system performance [4], [5]. From the perspectives of target detection in a dynamic environment, sensor network design (e.g., sensor placement, sensor selection, and the associated decision criteria) relies on the statistical characteristics of target behavior in the surveillance region. For temporally and spatially distributed events, a potential problem may arise due to large volumes of data with spurious background noise; this problem could be partially mitigated by data collection with a sparse set of sensors across the network.

A. Motivation

The problem, presented in this paper, is motivated from the perspective of information fusion over a local network of passive sensors, where the background environment interferes with sensor measurements. This situation could cause unstable boundaries of the statistical decision space during the training process [6]. In such cases, it is necessary to find a network statistic that remains invariant even with a time-varying background environment so that stable decision boundaries can be estimated. Applications of such locally-distributed sensor networks include unattended ground sensing, underwater target detection, and Internet-of-Things deployed for activity recognition; in all such applications, the use of passive (and inexpensive) sensors may result in environmental interference, which leads to classification errors for individual sensors. Nevertheless network performance could be significantly improved by including information from all sensors and inferring the collective undistorted information statistics. A major challenge here is that, besides high-frequency measurement noise, the spectral characteristics of environmental

2168-2267 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

and event signals are largely similar in both amplitude and time scale. Consequently, the underlying sensor fusion algorithms must be able to represent the characteristics across the network, which are preserved during the events of interest and which could be used to reliably detect the events of interest. Another challenge is that the statistics of the dynamic environment may be nonstationary and completely unknown. Therefore, sensor fusion algorithms for target detection must be executed in an unsupervised (or semi-supervised) fashion to adapt to environmental conditions.

B. Related Work

From the perspective of decision making, sensor fusion is important to reach a reliable decision for situation awareness. Most of the work in sensor network analysis is focused on model-based analysis, where performance is optimized for different performance criteria (e.g., target localization [7], target tracking [8], [9], sensor querying [10], and overall performance improvement [11]). Performance analysis from a data-driven perspective is challenging largely due to the lack of accurate models for data representation. Moreover, the problem becomes more complex when the environment and event dynamics are coupled and thus, the environment dynamics cannot be ignored for learning the data-driven statistical models for the events of interest. Additionally the development of parameterized models to account for environmental variation requires large amounts of training data that may not be easily accessible for many practical applications in sensor networks. For efficient operations, the sensor network must capture pertinent low-dimensional information across the network, which remains invariant, or may suffer very little distortion, due to changes in the environmental conditions. Recently, much work have been reported on information fusion in sensor networks under the data-driven paradigm with applications to target detection, intrusion detection, fault detection, denial-of-service attack, etc. [12]. A survey on information fusion in sensor networks could be found in [12] and [13] and that on data aggregation in [14]. Sensor fusion-based target detection has been extensively reported in the literature using detection theory-based local decisions [15], classification-based distributed detection [16], and fusion of local detection thresholds [17].

The work reported in this paper is a significant extension of the authors' earlier paper [18] that was presented in a conference as a preliminary version. The network data are first linearly decomposed into respective orthogonal components by using the standard principal component analysis (PCA) [19], and the network information state is then represented as the component that embeds the maximum variance in the network data. The time series at each sensor node is discretized into a symbol sequence for information compression. Eventually, two different types of Markov machine models are constructed; one based on the sensor symbol sequence itself [20], [21], and, in the other one, the individual sensor data are conditioned on the approximate network state [22]. Finally, the contribution of each sensor to the network information is computed based on the difference between the conditional entropy of these two types of Markov machines, which is inspired from the concept of transfer entropy [23]. For detecting the correlations across the network, the concept of transfer entropy is used, which is treated as a measure for causality detection [24], [25].

C. Contributions

While the earlier work by Li *et al.* [26] presented a feature-level fusion algorithm for passive sensor networks, this paper presents algorithms for analysis and improvement of information fusion across a sensor network by modeling the correlations between measurements of different sensors across the network using **x**-Markov machines [22]. The proposed method has been experimentally validated for multisensor target detection on a laboratory-scale network setting that was reported in a recent publication [26]. The results suggest that the proposed information-theoretic, conditional entropy-based sensor fusion algorithm is robust to the dynamic environment and it is able to achieve near-perfect performance for a small alphabet size (used for discretization of data).

Major contributions of this paper reported in this paper are outlined below.

1) *Development of an information-theoretic sensor fusion algorithm for target detection*, which is based on cross-dependencies between the network state and sensor states in a local network. The proposed algorithm is shown to be independent of the placement of the sensors and their individual identities.
2) *Experimental validation in a laboratory setting*, which demonstrates its efficacy for target detection. Dependence of the algorithm on the associated hyperparameters of the algorithm is illustrated using achievable target detection rates.
3) *Performance enhancement*, which is established by comparison of the proposed method with a recent feature-level sensor fusion method [26].

D. Organization

This paper is organized in five sections including the current section. Section II briefly describes the underlying principles of symbolic analysis that is used for feature extraction of the time-series signals from sensors. Section III elaborates the algorithm developed in this paper along with the list of pertinent assumptions. Section IV briefly presents the experimental procedure and the results of experimental validation of the proposed algorithm. Finally, this paper is summarized and concluded in Section V along with recommendations for future research.

II. Background

This section briefly describes the concept of symbolic time series analysis as well as the associated principles of information theory, upon which the work reported in this paper is constructed. Although the information in this section is available in standard literature and the authors' previous publications, it is presented here in a succinct and coherent fashion for completeness of the paper.

A. Symbolic Time Series Analysis for Markov Modeling

This section briefly describes the underlying concepts of D-Markov and **x**D-Markov machine models. For a detailed analysis and interpretation, interested readers are referred to earlier publications [20]–[22], [27]. In symbolic analysis of time-series data, continuous sensor data are mapped to a discrete set, thus generating a discrete symbol sequence. The dynamics of the continuous system are then studied in the symbolic space, which can be encoded in a finite-memory, finite-state probabilistic machine [20]. The dynamics of the symbol sequences are modeled as a probabilistic finite state automaton (PFSA), which is defined as follows.

Definition 1 (PFSA): A PFSA is a tuple $\mathcal{M} = (\mathcal{Q}, \mathcal{A}, \delta, \Pi)$ where

- $\mathcal{Q}$ finite set of states of the automata;
- $\mathcal{A}$ finite alphabet set of symbols $s \in \mathcal{A}$;
- $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ state transition function;
- $\Pi : \mathcal{Q} \times \mathcal{A} \to [0, 1]$ emission matrix of dimension $|Q| \times |A|$. The matrix $\Pi = [\pi_{ij}]$ is row stochastic such that π_{ij} is the probability of generating symbol $\mathcal{A}_j$ from state q_i.

For symbolic analysis of time-series data, a class of PFSAs, called the D-Markov machine, has been proposed [20] as a suboptimal but computationally efficient approach to encode the dynamics of symbol sequences as a finite state machine.

Definition 2 (D-Markov Machine [20], [21]): A D-Markov machine is a statistically stationary stochastic process $S = \cdots s_{-1}s_0s_1\cdots$ (modeled by a PFSA in which each state is represented by a finite history of D symbols), where the probability of occurrence of a new symbol depends only on the last D symbols, that is

$$\Pr(s_n \mid \cdots s_{n-D}\cdots s_{n-1}) = \Pr(s_n \mid s_{n-D}\cdots s_{n-1})$$

where the positive integer D is called the depth (or memory) of the Markov machine.

A D-Markov machine is thus a Dth-order Markov approximation of the discrete symbolic process. The assumption of a finite-length memory is reasonable for many stable and controlled engineering systems that usually tend to forget their initial conditions. The D-Markov machine is represented as a PFSA and states of this PFSA are words of length D or less. The state transitions are described by a sliding block code of memory D that could be state-dependent [21], [27] and, often for simplicity, are assumed to be uniform for all states [28]. In this way, the symbolic system is approximated as a finite-memory PFSA with a memory of length D.

The information content of the time-series is compressed as a PFSA by approximating the states by words of finite length from the symbol sequence. The PFSA induces a Markov chain of finite order, and the parameters of the Markov chain (e.g., the stochastic matrix) are estimated from data by following a maximum *a priori* probability (MAP) approach [21]. Once the parameters are estimated, they can be used for different machine learning applications (e.g., pattern matching and clustering) with underlying data sets.

Next the concept of **x**D-Markov machine is introduced and the underlying concept is also pedagogically illustrated in Fig. 1. The concept of **x**D-Markov machines is a generalization of the D-Markov machines which models the conditional temporal statistics between two stochastic processes.

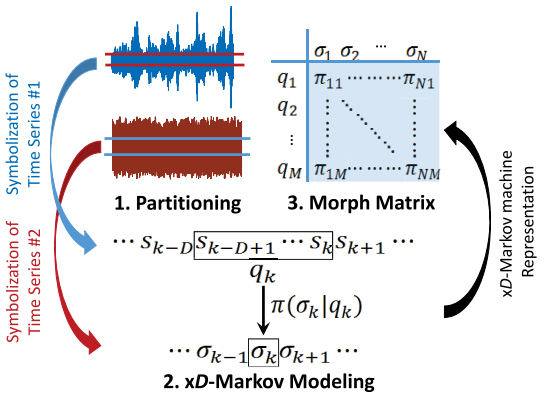


Fig. 1. Pedagogical representation of **x**D-Markov machines, where the (temporal) dynamics of a symbolic stochastic sequence are captured relative to another such sequence.

*Definition 3 (**x**D-Markov Machine) [22]:* Let $\mathbb{S}_1 = \{\ldots s_1s_2s_3\ldots\}$ and $\mathbb{S}_2 = \{\ldots\sigma_1\sigma_2\sigma_3\ldots\}$ be two symbol sequences. Then, a **x**D-Markov machine, where the Markov assumption holds for $\mathbb{S}_2$ with respect to the observations of $\mathbb{S}_1$, is defined as a 5-tuple $\mathcal{M}_{1\to2} \triangleq (\mathcal{Q}_1, \mathcal{A}_1, \mathcal{A}_2, \delta_1, \Pi_{12})$ such that:

1) $\mathcal{Q}_1 = \{q_1, q_2, \ldots, q_{|\mathcal{Q}_1|}\}$ is the state set corresponding to symbol sequence $\mathbb{S}_1$;
2) $\mathcal{A}_1 = \{s_1, \ldots, s_{|\mathcal{A}_1|}\}$ is the alphabet set of symbol sequence $\mathbb{S}_1$;
3) $\mathcal{A}_2 = \{\sigma_1, \ldots, \sigma_{|\mathcal{A}_2|}\}$ is the alphabet set of symbol sequence $\mathbb{S}_2$;
4) $\delta_1 : \mathcal{Q}_1 \times \mathcal{A}_1 \to \mathcal{Q}_1$ is the state transition mapping. It is noted that the PFSA structure is built on $\mathbb{S}_1$ and thus, the transition map explains the same symbol sequence; however, the Markov assumption holds for $\mathbb{S}_2$ on the states inferred in $\mathbb{S}_1$;
5) $\Pi_{12} : Q_1 \times \mathcal{A}_2 \to [0, 1]$ is the **x**-morph (pronounced as *cross*-morph) matrix of size $|\mathcal{Q}_1| \times |\mathcal{A}_2|$; the ijth element $\Pi_{12}(i, j)$ of Π_{12} denotes the probability of finding the symbol σ_j in the symbol string $\mathbb{S}_2$ at next time step while making a transition from the state q_i of the PFSA constructed from the symbol sequence $\mathbb{S}_2$.

Similarly, a 5-tuple $\mathcal{M}_{2\to1} \triangleq (\mathcal{Q}_2, \mathcal{A}_2, \mathcal{A}_1, \delta_2, \Pi_{21})$ is defined for $\mathbb{S}_1$ with respect to the observations of $\mathbb{S}_2$. One might be able to draw analogy with the conditional distribution of two random variables. However, the transition matrix of the cross-model captures also the temporal dependence between the two processes and thus, is more suitable for modeling of temporal processes.

B. Concepts of Entropy in Information Theory

This section very succinctly introduces a few standard concepts of information theory, which will be used later for the analysis presented in this paper. Interested readers are referred to [23] for an introductory text in information theory and entropy.

Definition 4 (Mutual Information [23]): Formally, the mutual information of two discrete random variables, X and Y, is defined as

$$I(X;Y) = \sum_{\substack{x\in X,\\ y\in Y}} p_{XY}(x,y)\ \log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right). \tag{1}$$

The mutual information is a measure of the relative dependence between two random variables and it is a symmetric function of X and Y, which can be equivalently expressed as

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

$$= H(Y) - H(Y|X). \tag{3}$$

Definition 5 (Transfer Entropy [24], [25]): Transfer entropy from a (discrete) random process X_t to a (discrete) random process Y_t ($t \in \mathbb{N}$) is measured by the mutual information between Y_t and the history of the random variable $X_{t-D}^{t-1} = [x_{t-D}\dots x_{t-1}]$ given the history of the random variable $Y_{t-D}^{t-1} = [y_{t-D}\dots y_{t-1}]$ in the condition

$$T_{X\to Y} = I\left(Y_t; X_{t-D}^{t-1} | Y_{t-D}^{t-1}\right)$$

$$= H\left(Y_t|Y_{t-D}^{t-1}\right) - H\left(Y_t|X_{t-D}^{t-1}, Y_{t-D}^{t-1}\right). \tag{4}$$

Given the past information on X and Y, transfer entropy quantifies the reduction of uncertainty in future values of Y. It is a measure of directed information transfer between two random processes. This idea of transfer entropy is closely related to the information contribution measure introduced later in this paper.

III. Technical Approach

This section presents the proposed method of measuring the information contribution of each sensor in the network under a dynamic environment. First, a general sensor model is formulated and the basic assumptions are provided with due consideration of feasible applications. Then, the proposed measurement model is developed based upon the background material provided in Section II.

A. Modeling and Assumptions

Let a sensor network consist of N sensor nodes for surveillance in the region of interest. The ith sensor node generates a real-valued time series $\mathbf{y}_i = \{y_i[1], y_i[2], \dots, y_i[n_{\text{obs}}]\}$ for $n_{\text{obs}} \in \mathbb{N}$. The sensor signal $y_i[k]$ at an instant k is a (nonlinear) function of environment state (which is dynamic) $S_E[k]$, target state $S_T[k]$, and measurement noise $v[k]$. No assumptions are made on the underlying information about the environment and target state distribution and thus, the problem is entirely dynamic data-driven. The respective sensor signals for both "target absent" (i.e., environmental disturbances only) and "target present" cases at time $k = 0, 1, \dots, n_{\text{obs}}$ could be modeled by two distinct joint probability distributions.

In this paper, the information state of the sensor network is represented by the first principal component $\mathbf{x}$ obtained from the matrix of all of the sensors' time series $\mathbf{Y} = [\mathbf{y}_1\mathbf{y}_2\dots\mathbf{y}_N]$. Let v_i be the normalized eigenvectors of the real symmetric matrix $\mathbf{YY}^T$ corresponding to the (real positive) eigenvalues λ_i that are arranged in decreasing order of magnitude, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$. The corresponding (normalized) eigenvectors u_i in the original data space are obtained in terms of v_i and λ_i as follows [19]:

$$u_i = \frac{v_i}{\sqrt{N\lambda_i}}\mathbf{Y}^T,\ i = 1, 2, \dots, N. \tag{5}$$

Then, the first (i.e., largest) principal component $\mathbf{x}$ of all sensor time series matrix can be expressed as

$$\mathbf{x} = \mathbf{Y}u_1. \tag{6}$$

It is noted that the first principal component $\mathbf{x}$ is a weighted linear combination of the readings, and it contains both environment and target information under the target present cases. To simplify the problem in the absence of any target or environment information, the following assumptions are made on the local sensing model.

1) The sensors are passive and homogeneous, i.e., all sensors are of the same modality.
2) Sensor deployment is heterogeneous, i.e., sensors are not necessarily co-located and their orientations may be different.
3) Within a given time span of observation, there is at most one (moving) target.
4) The time scale of environmental changes is largely similar to that of event occurrence (e.g., target motion).

The usage of homogeneous sensors is cost-effective and placing them non-co-located with different orientations provides heterogeneity that allows a variety of views of the target and environment. Therefore, different sensors may yield different measurements even in the presence of the same single target. The sensors are usually looking for rare events (i.e., hard-to-find targets), which generally means only one target, will appear in a field-of-view within a short time span, or else the field of view would be too large to detect a faint target. In other words, the observed time series outputs of the sensors correspond to the same event (i.e., the same target if it is present). Finally, the real-world environment is often dynamically changing during the observation of encountering hard-to-find targets. Thus, the above assumptions are consequences of phenomena that may naturally occur at a node of a sensor network in real-life applications.

B. x*D*-Markov Machines

Following Definition 3, a x*D*-Markov machine represents a Markov model that incorporates the behavior of a symbolic stochastic process based on the observations of another stochastic process by using the algebraic structure of a finite state automaton (FSA). The concept is qualitatively elucidated in Fig. 1 (please note that the figure has been simplified for pedagogical understanding), where the states of the x*D*-Markov machine $\mathcal{M}_{1\to 2}$ are created as words from the symbol string $\mathbb{S}_1$ and the emission probabilities of symbols are inferred from the symbol string $\mathbb{S}_2$. It is noted that the x*D*-Markov machine $\mathcal{M}_{2\to 1}$ could be similarly constructed. A step-by-step procedure to generate a x*D*-Markov machine from a sensor network time series is described next.

1) Signal Preprocessing: First, the raw time series for each sensor is normalized to be zero-mean and unit-variance as

$$\bar{y}[k] = \frac{y[k] - \mu_y}{\sigma_y} \text{ for } i = 1, 2, \ldots, L \tag{7}$$

where μ_y is the mean and σ_y is the standard deviation of the raw time series y having a finite length of L. The objective of this signal preprocessing is to remove the undesirable effects (e.g., due to bias and spurious noise) from the time series before symbolization. As an additional preprocessing step, independent identically distributed zero-mean Gaussian noise ($\xi \sim \mathcal{N}(0, \sigma_\xi^2)$) is added to the time series uniformly at each data point. The preprocessed signal for the ith sensor at each time instant k is expressed as

$$\hat{y}_i[k] = \bar{y}_i[k] + \xi. \tag{8}$$

The addition of independent identically distributed noise to signals before discretization is motivated by the fact that adding noise to inputs acts as regularization in machine learning techniques [29]. During the symbolization process, the addition of noise has been empirically shown to improve the signal-to-noise ratio [30]. To avoid loss of significance of the embedded information, which could undermine the PFSA feature, the standard deviation of the injected Gaussian noise is chosen as $\sigma_\xi^2 = 1/|\mathcal{A}|$ in this paper, where $|\mathcal{A}|$ is the alphabet size of the underlying FSA (see Definition 1).

2) Symbolization of Time Series: This step requires partitioning (also known as quantization) of the time series data of the measured signal. The signal space is partitioned into a finite number of cells that are labeled as symbols, i.e., the number of cells is identically equal to the cardinality $|\mathcal{A}|$ of the (symbol) alphabet $\mathcal{A}$. The choice of alphabet size $|\mathcal{A}|$ largely depends on the specific data set and the allowable loss of information (e.g., leading to error of detection and classification) [27], [31]. However, in general, the alphabet size is selected using cross-validation. The partitioning depends largely on the underlying statistics of the data—the probability distribution of the data and the corresponding temporal statistics of the discrete data. While it is difficult to provide a universal rule for partitioning, for supervised learning problems, cross-validation provides a reasonable solution. Interested readers are referred to [32] for a review of related techniques. For unsupervised learning tasks, the problem remains largely open.

The ensemble of time series data are partitioned by using a partitioning tool (e.g., maximum entropy partitioning (MEP) [21]) that maximizes the entropy of the generated symbols; therefore, the information-rich cells of a data set are partitioned finer and those with sparse information are partitioned coarser. As an example for the 1-D time series in Fig. 1, the alphabet $\mathcal{A}_1 = \{s_1, s_2, s_3\}$, i.e., $|\mathcal{A}| = 3$, and two horizontal lines divide the ordinate (i.e., y-axis) of the time series profile into three mutually exclusive and exhaustive regions. These disjoint regions form a partition, where each region is labeled with one symbol from the alphabet $\mathcal{A}_1$.

*3) **x**D-Markov Modeling:* The problem of **x**D-Markov modeling could be stated as follows (see Definition 3). Given two stochastic symbolic processes $\mathbb{S}_1$ and $\mathbb{S}_2$, the task is to create a generative model $\mathcal{M}_{1\to 2}$ for $\mathbb{S}_2$ based on the observations from $\mathbb{S}_1$. This requires an inference of the generative model of causal dependence between the two processes $\mathbb{S}_1$ and $\mathbb{S}_2$. In this model, the temporal dependence is captured by assuming a Markov structure between the observed variables. Such a cross-dependence is represented as **x**-automata (pronounced as *cross* automata), which induces a Markov chain.

The states of the **x**D-Markov machine are inferred by using a state-splitting algorithm that, under the D-Markov assumption, leads to states of variable depth for the **x**D-Markov machine. In particular, the states of the **x**D-Markov machine are split by using conditional entropy as the metric, where the largest decrease in conditional entropy is used to select the state to be split. The conditional entropy of a **x**D-Markov machine is defined next.

*Definition 6 (Conditional Entropy of **x**D-Markov Machines [22]):* The conditional entropy of a **x**D-Markov machine $\mathcal{M}_{1\to 2} = (\mathcal{Q}_1, \mathcal{A}_1, \mathcal{A}_2, \delta_1, \Pi_{12})$ representing the causal dependence of the stochastic symbolic process $\mathbb{S}_2 = \{\sigma_t \in \mathcal{A}_2 : t \in \mathbb{N}\}$ on the stochastic symbolic process $\mathbb{S}_1 = \{s_t \in \mathcal{A}_1 : t \in \mathbb{N}\}$ is defined as

$$\begin{aligned} H(\mathcal{A}_2|\mathcal{Q}_1) &\triangleq \sum_{q_i \in \mathcal{Q}_1} P(q_i) H(\mathcal{A}_2|q_i) \\ &= -\sum_{q_i \in \mathcal{Q}_1} \sum_{\sigma_j \in \mathcal{A}_2} P(q_i)\Pi_{12}(i,j) \log \Pi_{12}(i,j) \end{aligned} \tag{9}$$

where $P(q_i)$ is the probability of the **x**D-Markov machine state $q_i \in \mathcal{Q}_1$ and $P(s_j|q_i)$ is the conditional probability of the symbol $s_j \in \mathcal{A}_2$ given that a **x**D-Markov machine state $q_i \in \mathcal{Q}_1$ has been observed.

The process of splitting a state $q \in Q_1$ of the **x**D-Markov machine $\mathcal{M}_{1\to 2} = (\mathcal{Q}_1, \mathcal{A}_1, \mathcal{A}_2, \delta_1, \Pi_{12})$ is executed by replacing the symbol block for q by its branches given by the set $\{sq : s \in \mathcal{A}_1\}$. Then, the maximum reduction in the conditional entropy of $\mathcal{M}_{1\to 2}$ is the governing criterion for selecting the state to be split, based on the user-input parameters of maximum number of states $n_{\max}$ or threshold η_{spl} [21]. As a result, not all the states are split and thus, this creates a variable-depth structure of $\mathcal{M}_{1\to 2}$. The underlying state splitting algorithm is delineated in Algorithm 1.

Remark 1: Algorithm 1 creates a model with suboptimal predictive accuracy; however, this is helpful in restricting the state-space size of the constructed automaton. Another point to be noted is that the entropy rate is a submodular function of the size of state-space; thus adding more states leads to a decreased rate of reduction in entropy rate. Hence, a stopping rule is used to terminate the splitting algorithm based on either the rate of change of entropy rate or the maximum number of states allowed in the final automaton. It is further noted that this allows the flexibility of inferring states with different memories, depending on the data statistics. Both the number of states $n_{\max}$ and η_{spl} are chosen during training using cross-validation of modeling or classification accuracies. It is noted that the transition model $\delta(q, s)$ is estimated based on the topology of the FSA with a known memory.

Algorithm 1 State Splitting for Variable-Depth **x**D-Markov Machine

Input: Symbol sequences $\mathbb{S}_1 = \{\cdots s_1 s_2 s_3 \cdots : s_i \in \mathcal{A}_1\}$ and $\mathbb{S}_2 = \{\ldots \sigma_1 \sigma_2 \sigma_2 \cdots : \sigma_i \in \mathcal{A}_2\}$.
User Input: Maximum number of states n_{max} or threshold η_{spl}
Output: **x***D*-Markov $\mathcal{M}_{1\to 2} = (\mathcal{Q}_1, \mathcal{A}_1, \mathcal{A}_2, \delta_1, \Pi_{12})$
Initialize: Create a 1-Markov machine $Q^* \coloneqq \mathcal{A}_1$
repeat
  $Q_1 \coloneqq Q^*$
  $Q^* = \arg\min_{Q'} H(\mathcal{A}_2|\mathcal{Q}')$
  where $Q' = Q \backslash q \cup \{sq : s \in \mathcal{A}_1\}$ and $q \in Q_1$
until $\left(|Q^*| \le n_{max}\right)$ or $\left(H(\mathcal{A}_2|Q_1) - H(\mathcal{A}_2|Q^*) \le \eta_{spl}\right)$
for all $q \in Q^*$ and $s \in \mathcal{A}_1$ **do**
  if $\delta(q, s)$ is not unique **then**
    $Q^* \coloneqq Q^* \backslash q \cup \{sq : s \in \mathcal{A}_1\}$
    /** Consistent algebraic structure of $\mathcal{M}_{1\to 2}$ **/
  end if
end for
return $\mathcal{M}_{1\to 2} = (\mathcal{Q}_1, \mathcal{A}_1, \mathcal{A}_2, \delta_1, \Pi_{12})$

Remark 2: The hyperparameters associated with the **x***D*-Markov modeling are the partitions of the individual time-series and the depth (memory) of the corresponding Markov model. To reduce the complexity of the associated models, the search of the hyperparameters is generally done by fixing a partitioning (e.g., MEP as described earlier) and then an estimate of the depth (memory) of the corresponding Markov model could be obtained from the behavior of the cross entropy rate of the Markov model, as was illustrated in [22]. However, much more generalized approaches (e.g., minimum description length [33] or even cross-validation) can be adopted to estimate the optimal (in some desired sense) hyperparameter. In this paper, a model with a variable depth (memory) is selected by state-splitting with entropy rate [21] as outlined in Algorithm 1. The state-splitting algorithm is terminated after the model exceeds a number of states or from observing the entropy-rate behavior (i.e., the algorithm is stopped when the rate of entropy-rate decrease is smaller than a threshold). The worst-case time complexity of the state-splitting-based tree search is $O(\log(N))$ where N is the maximum number of states (equivalent to greedy nearest-neighbor tree search). The worst-case time complexity of the stochastic matrix estimation is $O(|\mathcal{A}| \times |\mathcal{Q}| + |\mathbb{S}|)$.

Once the hyperparameters (i.e., alphabet size $|\mathcal{A}|$ and depth D) are fixed, the statistical parameters of the **x***D*-Markov model are estimated by using the maximum *a posteriori* (MAP) rule with uniform prior [21]. Let $n_{\text{count}}(\sigma_j|q_i)$ denote the number of times that a symbol σ_j is generated in $\mathbb{S}_2$ when the state q_i as the symbol string is observed in $\mathbb{S}_1$. The MAP estimate of the emission probability of the symbols $\sigma_j \in \mathcal{A}_2$ conditioned on $q_i \in \mathcal{Q}_1$ is estimated by frequency counting [21] as follows:

$$\hat{p}(\sigma_j|q_i) = \frac{1 + n_{\text{count}}(\sigma_j|q_i)}{|\mathcal{A}_2| + \sum_{\ell=1}^{|\mathcal{A}_2|} n_{\text{count}}(\sigma_\ell|q_i)}. \tag{10}$$

If no event is generated at a combination of symbol $\sigma_j \in \mathcal{A}_2$ and state $q_i \in Q_1$, then there should be no preference to any particular symbol and thus, we choose $\hat{p}(\sigma_j|q_i) = (1/|\mathcal{A}_2|)$. The above procedure guarantees that the **x***D*-Markov machines, constructed from two (finite-length) symbol strings, must have an (elementwise) strictly positive **x**-morph matrix Π_{12}.

C. Information Contribution Measures in Sensor Network

This section introduces an information contribution measure of the individual sensor time series $\mathbf{y}_i$ ($i \in [1, 2, \ldots, N]$) conditioned on the first principal component in the time series $\mathbf{x}$, as being a representative of the overall information from the sensor network. The objective here is to evaluate the contribution of each sensor to the information content of the network. All time series are preprocessed and discretized into symbol sequences for reduction of communication and computation costs before initiating any network operation.

Following Section II-B, the transfer entropy from the time series $\mathbf{x}$ to an individual time series $\mathbf{y}_i$ is expressed as

$$\begin{aligned} T_{\mathbf{x}\to\mathbf{y}_i} = {} & H(\mathbf{y}_i[k] \mid \mathbf{y}_i[k-D:k-1]) \\ & - H(\mathbf{y}_i[k] \mid \mathbf{x}[k-D:k-1], \mathbf{y}_i[k-D:k-1]) \end{aligned} \tag{11}$$

which quantifies the amount that the uncertainty of predicting future the value $\mathbf{y}_i[k]$ is (possibly) reduced as a consequence of having the additional information $\mathbf{x}_i[k-D:k-1]$. In this way, a new information measure is introduced, which compares the predictability of the future $\mathbf{y}_i[k]$ given the past value of itself $\mathbf{y}_i[k-D:k-1]$ and the past value of the first principal component $\mathbf{x}[k-D:k-1]$ with a predefined depth (memory) D. This new information measure is expressed as

$$\begin{aligned} \Delta I_{\mathbf{x}\to\mathbf{y}_i} = {} & H(\mathbf{y}_i[k] \mid \mathbf{y}_i[k-D:k-1]) \\ & - H(\mathbf{y}_i[k] \mid \mathbf{x}[k-D:k-1]). \end{aligned} \tag{12}$$

It is noted that $\Delta I_{\mathbf{x}\to\mathbf{y}_i}$ is an information contribution measure between $\mathbf{x}$ and $\mathbf{y}_i$. If $\Delta I_{\mathbf{x}\to\mathbf{y}_i} < 0$, then the conditional entropy of the future $\mathbf{y}_i[k]$ given the past $\mathbf{x}[k-D:k-1]$ is higher than that given the past $\mathbf{y}_i[k-D:k-1]$ only. In that case, the predictability of $\mathbf{y}_i[k]$ by observing $\mathbf{x}[k-D:k-1]$ is less than that of $\mathbf{y}_i[k-D:k-1]$. Since $\mathbf{x}$ is assumed to represent the overall information in the network, the sensor time series $\mathbf{y}_i$ should influence $\mathbf{x}$ in time; the rationale is that $\mathbf{x}$ contains information about the network. By a similar argument, the first principal component $\mathbf{x}$ influences the sensor time series $\mathbf{y}_i$, because the information derived from sensor data is commonly shared in the network when $\Delta I_{\mathbf{x}\to\mathbf{y}_i} > 0$. Equivalently, sensors with a detected target are expected to have a negative information contribution measure, while that of sensors that capture environment information only is likely to be positive.

To compute the proposed information measure in the framework of **x***D*-Markov machines, the set of states Q is assigned as all words (i.e., symbol blocks) w having the same length as the depth D, i.e., $|w| = D$. Equivalently

$$Q = \bigcup_{|w|=D} w. \tag{13}$$

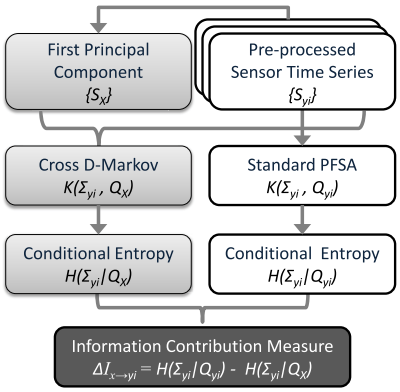


Fig. 2. Flow chart to generate information contribution for each sensor in the network.

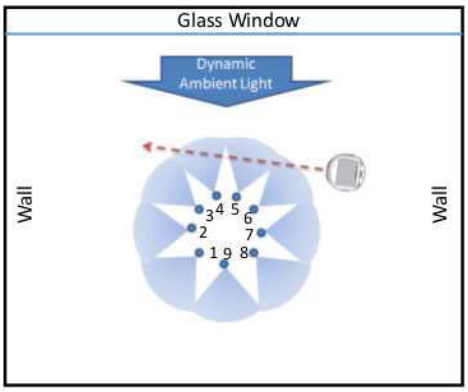


Fig. 3. Schematic of the experimental setup for target detection.

With the conditional entropy of $\mathbf{x}D$-Markov machine defined in (9), the proposed information measure is reformulated as

$$\Delta I_{\mathbf{x}\to\mathbf{y}_i} = H\big(\mathcal{A}_{\mathbf{y}_i}|Q_{\mathbf{y}_i}\big) - H\big(\mathcal{A}_{\mathbf{y}_i}|Q_{\mathbf{x}}\big). \tag{14}$$

Fig. 2 presents the flow chart of the procedure to generate the information contribution measure for each sensor in the sensor network.

IV. Experimental Validation

This section presents the results of experimental validation on a laboratory apparatus that is served by the sensor network. Fig. 3 depicts the layout of the laboratory apparatus; although this figure was presented in a recent publication [6], [26], it is replicated in this paper for convenience of referring to the experimental results.

The sensor network is configured as a ring of nine TCRT5000 infrared sensors; details of the collected time-series data characteristics could be found in [6]. A computer-instrumented and computer-controlled Khepera III mobile robot [34] serves as a single moving target. The dynamic environment and the associated environmental disturbances are emulated as variations in the daylight intensities on partially cloudy days. As seen in Fig. 3, the local area of the sensor network is placed at the center of a square room in which only one wall has open windows that are exposed to the sun. For a more detailed description of the laboratory facility, the reader is referred to the website http://nrsl.mne.psu.edu of Networked Robotic Systems Laboratory at Penn State. All experiments were conducted during days under partially cloudy conditions, where the sunlight is intermittently blocked by clouds. During the experiments, the moving target travels at a constant speed in straight lines between the sensor network and the ambient light source, as illustrated in Fig. 3. The infrared sensors are oriented toward the moving target and are subjected to disturbances when the target is moving in and out, which causes intermittent blocking of the ambient light source. Due to the orientation of sensors, effects of the environment are different for different individual sensors. Sensors that face the windows (i.e., the ambient light source) have different levels of reading when compared with the rest of the sensors. When a target is moving in, ambient light sources are partially blocked for some of the sensors, which increases their readings temporarily. Changes in the ambient light affect the moving target as the readings of some of the sensors fluctuate more significantly. Thus, it becomes difficult to correctly detect a target by simply estimating a threshold on the sensor readings under environmental changes [26].

In total, 50 experiments with environment-only (i.e., absence of a target) and another 50 experiments with target present have been conducted. For each experiment, each of the nine sensors record synchronized data for about 65 s with an average sampling rate of ∼18.5 Hz.

A. Signal Preprocessing for Robustness

Fig. 4 presents the results of noise injection on a target absent (i.e., dynamic environment only) event. Once symbol strings are obtained from the sensor time series via MEP, PFSA features are generated accordingly via the algorithm described in Section II-A. The distance between two PFSA features from different sensors is obtained by the cosine distance function $D_C(\bullet, \bullet)$, the rationale of which is explained in an earlier publication [26]

$$D_C(\mathbf{a}, \mathbf{b}) = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} {a_i}^2}\sqrt{\sum_{i=1}^{n} {b_i}^2}} \tag{15}$$

where $\mathbf{a} = [a_1 a_2 \cdots a_n]$ and $\mathbf{b} = [b_1 b_2 \cdots b_n]$ are two real vectors with the same dimension $n \in \mathbb{N}$.

The objective of preprocessing of time series is to increase the similarity of PFSA features among sensors for the target absent case, while maintaining the dissimilarity of PFSA features between sensors with target detected and the others under target present (see Section III-B1 for intuition and more details of noise addition). Table I presents the average ratio of top cluster divergence for target present to target absent cases with/without adding noise for different choices of the alphabet size for PFSA feature generation. The sensor feature cluster divergence for the target present cases (with noise injection in the time series) is amplified compared to the measures under the target present cases. This indicates that the performance of using the PFSA feature to distinguish target present from target absent is improved by signal preprocessing.

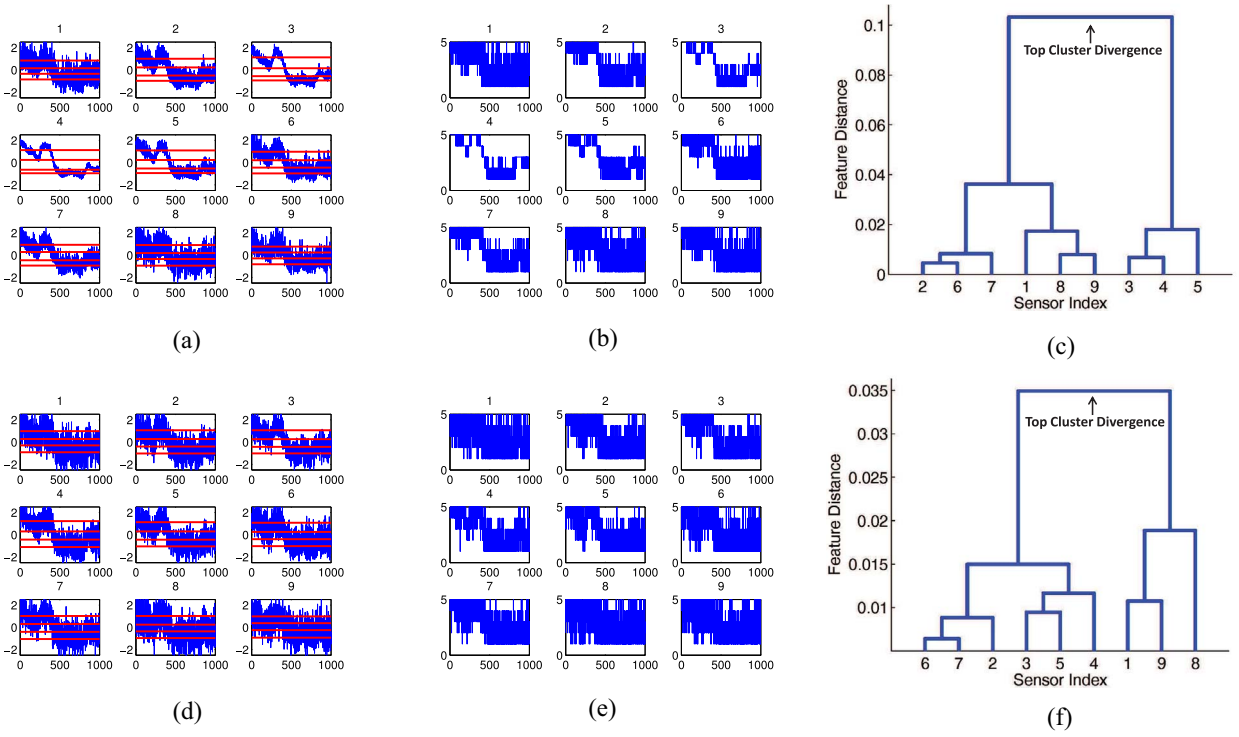


Fig. 4. Feature clustering with alphabet size $|\mathcal{A}| = 5$ and cosine distance measurement used as the distance among PFSA features. Plates (a)–(c), respectively, show partitioning, symbolization, and feature clustering before noise injection. Plates (d)–(f), respectively, show the same after noise injection.

TABLE I
AVERAGE FEATURE DIVERGENCE RATIO BETWEEN TARGET PRESENT AND TARGET ABSENT CASES

Alphabet Size		2	3	4	5	6	7	8
Feature Divergence Ratio: Target Present / Target Absent	without noise	2.42	2.06	2.03	2.05	2.03	1.98	1.96
	with noise	3.28	3.90	4.33	4.33	4.34	3.88	3.58

B. Information Contribution Measures

As stated earlier, the sign of the information contribution measure reflects the direction of information flow between a pair of time series. A positive value of information contribution indicates that the first principal component, which represents overall network information, dominates the sensor under consideration; a negative value indicates that the sensor has its own unique information (i.e., transition pattern in time series) that leads to a contribution to the network information content.

The proposed information contribution measure is computed as the difference of the conditional entropy between two types of Markov machines constructed from sensor time series. The conditional entropy of a standard (self) D-Markov machine for each sensor time series represents the uncertainty of predicting the future based on its own past. The conditional entropy of $\mathbf{x}D$-Markov machines between sensor time series and the first principal component measures the uncertainty of predicting the future sensor time series based on the first principal component.

Fig. 5 depicts the average values of conditional entropies of both standard and $\mathbf{x}D$-Markov machines for both target absent and target present cases under the alphabet size of $|\mathcal{A}| = 4$. The conditional entropy of the standard D-Markov machine for target absent is larger than that of $\mathbf{x}D$-Markov machine, possibly except for sensor 3 that dominates the information contribution relative to the first principal component due to its location and orientation to the ambient light. On the other hand, sensors 3–5 have significantly smaller conditional entropies for standard D-Markov machines as compared to $\mathbf{x}D$-Markov machines. Since sensors 3–5 are the three most likely sensors that may have detected the target, their contributions to the first principal component are more significant as compared to other sensors that may only contain the environmental information.

Fig. 6 depicts the average values of proposed information contribution measures of each sensor under both target present and target absent cases. For target absent, the information contribution measures are all positive and similar among all sensors, except for sensor 3. On the other hand, the information contribution measures for target present have distinguishable patterns for sensors. It is noted that sensors 3–5, which would most likely detect the target, have negative average information contribution measures, while the sensors with no target information yield positive values of small magnitude.

C. Alternative Network Information Representation

In this paper, the first principal component is chosen as the representation of overall network information, since it

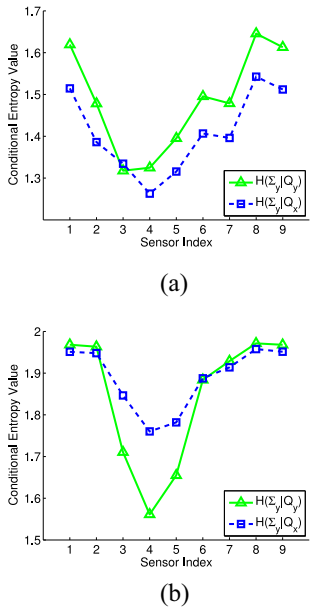


Fig. 5. Average conditional entropies for (a) target absent and (b) target present.

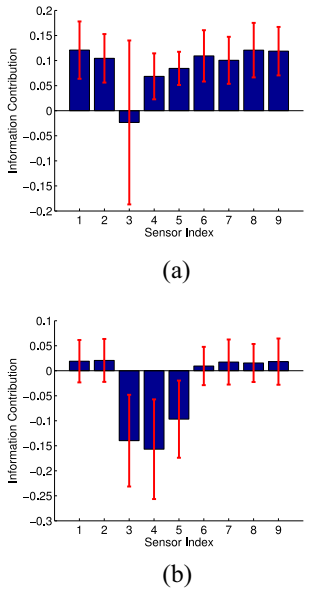


Fig. 6. Mean (i.e., blue bar) and standard deviation (i.e., red line) of information contribution measures for (a) target absent and (b) target present cases.

maximizes the variance of the projection from sensor quantized symbol strings. Following (5) and (6), the first principal component is obtained as a weighted linear combination of the original sensor strings.

In general, the first principal component, as a representation of the overall network information, is a linear combination of sensor symbol sequences with case-dependent numerical weights. To simplify the step of acquiring network information, an alternative representation of information is proposed based on data averaging (DA)

$$X = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i. \tag{16}$$

The above information representation is a uniformly weighted linear combination of all sensor symbol sequences, which incurs a much lower computational cost, compared to the first principal component. As shown in the next section, this simplification provides an equivalent target detection performance to that of the first principal component.

D. Results of Target Detection

In a previous publication [26], the decision rule for target detection in a dynamic environment had been constructed by thresholding of the clustering divergence of standard Markov machine features, constructed from time series without any noise injection. It showed (nearly) perfect performance with a relatively large alphabet size (e.g., $|\mathcal{A}| = 10$) and the cosine distance measure between features. In this paper, the results of feature clustering algorithm are compared with and without injection of independent and identically distributed zero-mean Gaussian noise, as well as with the result of the proposed information contribution measures of sensors. As the conditional entropy of the Markov models depends on the hyperparameters (i.e., alphabet size $|\mathcal{A}|$ and depth D), they are kept the same for both the target present and target absent cases for a consistent comparison. However, the effects of changing these hyperparameters are shown later in this section.

In this paper, the decisions on target detection are made by thresholding on the selected information contribution measure among all available sensors in the network for each testing event. It is seen in Fig. 6 that only sensor 3 has a negative mean value of information contribution measure for the target absent cases, along with a significant standard deviation. On the other hand, there are three sensors with negative mean information contribution measure and relatively small standard deviations for the target present cases. So, the second to the last largest information contribution measures of all nine sensors is chosen as the criterion for the target detection for performance robustness in this paper.

To determine the threshold based on the information causality measure values, the target detection problem is formulated as a binary hypothesis test in terms of the hypothesis pair as

$$\begin{cases} H_0 : X \sim \mathcal{P}_0 \\ H_1 : X \sim \mathcal{P}_1 \end{cases} \tag{17}$$

where X represents the information contribution measure value from the sensor of interest; and the probability measures $\mathcal{P}_0$ and $\mathcal{P}_1$ represent the feature distance under the null hypothesis H_0 (i.e., target absent) and the alternate hypotheses H_1 (i.e., target present), respectively. The decision threshold η, which is user-selectable, yields the following decision logic:

$$X \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{18}$$

It is a usual practice to select η from the receiver operating characteristic (ROC) curve [35] that is constructed from X computed from all training events of both hypotheses H_0 and H_1.

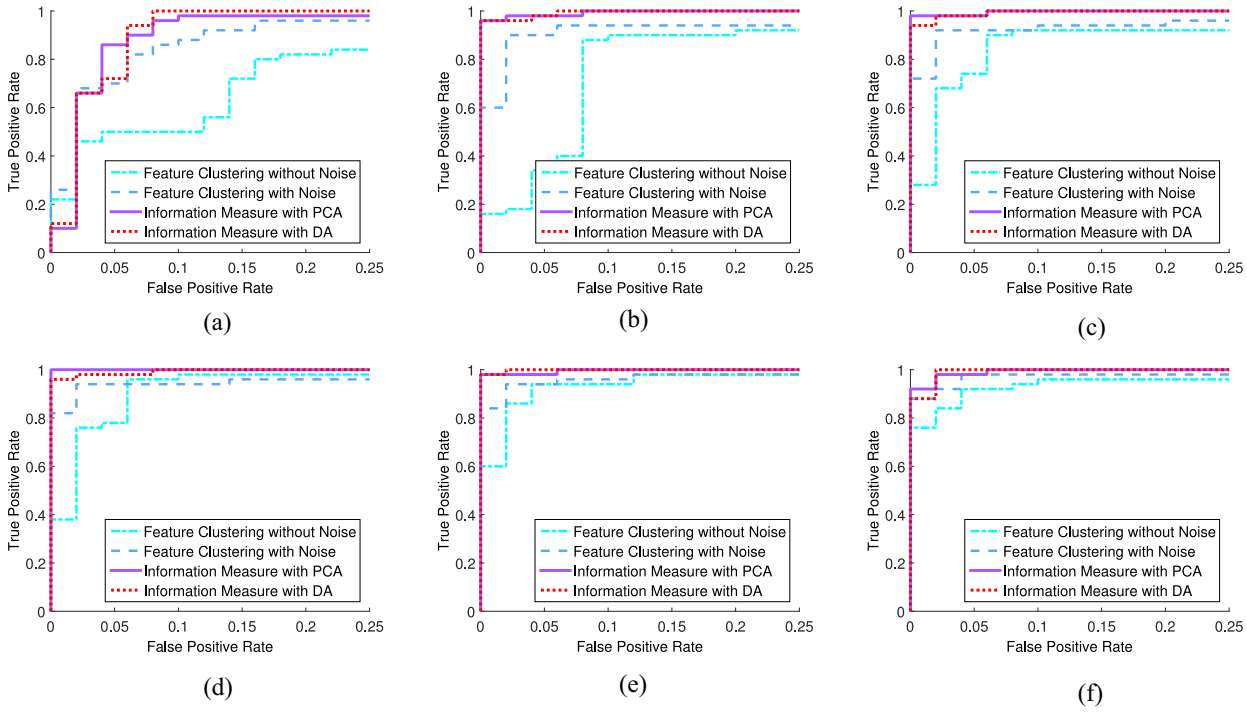


Fig. 7. ROC curves for target detection under different choices of the alphabet size. (a) $|\mathcal{A}| = 3$. (b) $|\mathcal{A}| = 4$. (c) $|\mathcal{A}| = 5$. (d) $|\mathcal{A}| = 6$. (e) $|\mathcal{A}| = 7$. (f) $|\mathcal{A}| = 8$.

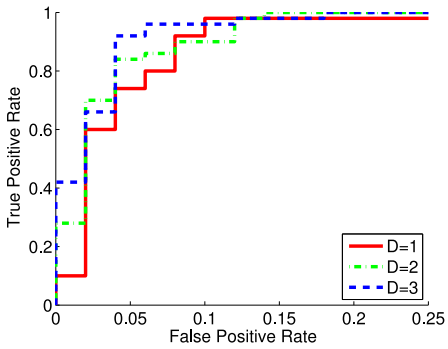


Fig. 8. ROC curve of target detection for $|\mathcal{A}| = 3$ with different depths.

Fig. 7 presents the ROC curves for target detection for different alphabet sizes ranging from $|\mathcal{A}| = 3$ to $|\mathcal{A}| = 8$ under the following four different algorithms.

1) Feature clustering without noise injection.
2) Feature clustering with noise injection.
3) Information measure with PCA.
4) Information measure with (uniformly weighted) DA.

The results in Fig. 7 show that the performance of the ROC curves follows a general trend of improvement as the alphabet size increases. In particular, the clustering algorithm with injection of zero-mean Gaussian noise is generally superior to that with no noise injection. Apparently, the average performance of the proposed information contribution measure with PCA is the best among the four algorithms; unlike DA, the PCA algorithm can be used with a relatively smaller subset of available data. Furthermore, the PCA algorithm yields good target detection even with a choice of small alphabet size and it continues to obtain consistently good results as the alphabet size is increased. In Fig. 7, the rationale for slight performance degradation with a larger alphabet size (e.g., $|\mathcal{A}| > 6$) is possibly due to the finite data size and probability estimation by frequency counting (over-fitting with limited data). For a sufficiently large data size, the detection performance in terms of the ROC curves should be monotonically increasing with alphabet size $|\mathcal{A}|$.

Fig. 8 presents the ROC curves of target detection generated by the proposed information fusion method when the alphabet size $|\mathcal{A}| = 3$ and the temporal memory (or depth) is chosen at $D = 1, 2,$ and 3. The results show that, on the average, the target detection performance is noticeably improved as D is increased for $\mathbf{x}D$-Markov modeling. In particular, the true positive rate is significantly larger with increased D at small false positive rates in the range of 0.05–0.10. This observation indicates that more robust and accurate information contribution measures can be achieved if more detailed $\mathbf{x}D$-Markov machines are constructed. Therefore, the proposed algorithm appears to be able to consistently outperform the earlier algorithm [26] by using a smaller alphabet size, which reduces the communication load across the network in addition to reduced computation at individual sensor nodes. Furthermore, the proposed algorithm is not susceptible to the choice of a metric for measuring distances between the features, which played a central role in the earlier algorithm [26].

V. Conclusion

This paper develops a generalized framework and presents computation of the measure of information contribution in

passive sensor networks for target detection. Two different concepts have been presented for network information state representation: one based on simple averaging and the other based on eigenvalue decomposition, both of which have been experimentally validated for event detection in a dynamic environment on a laboratory-scale apparatus that serves as the local area of a sensor network. We have compared the work with a feature-level information fusion algorithm and have shown that we can outperform the same with very small computational burden (using compact models for small alphabet size). Also, the proposed fusion algorithm does not require the exact positioning of the sensors or their spatial correlation (assuming they remain the same during training and testing events). The algorithm could be used to detect events which are observed by a local network (i.e., sensors which observe an event at almost the same time when compared to the event dynamics).

While there are many issues that need to resolved by further theoretical and experimental research, the authors suggest the following potential topics of future research.

1) *Selection of Hyperparameters for Cross Machines:* An important future work is to optimize the hyperparameters of cross-machines, i.e., the right partition size and the related depth—a theoretical analysis for consistency of the technique would help establish the correctness of the proposed approach.
2) *Extended Experimental Validation:* The test apparatus should be expanded to accommodate large-scale sensor networks, because a larger sensor network would be able to characterize the performance of the proposed concepts in more complex settings (e.g., multiple targets and different target types).
3) *Comparison With Other Sensor Fusion Techniques:* While in this paper, the proposed concepts have been validated in the setting of the authors' earlier work for information fusion, an important future work is to compare the concepts with other existing techniques for data aggregation, which are available in open literature.
4) *Investigation of Network Information State:* This topic requires further investigation as the choice of network information state is expected to be largely system-specific.

REFERENCES

[1] I. D. Schizas and V. Maroulas, "Dynamic data driven sensor network selection and tracking," *Procedia Comput. Sci.*, vol. 51, no. 1, pp. 2583–2592, 2015.

[2] D. Bajović, B. Sinopoli, and J. Xavier, "Sensor selection for event detection in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4938–4953, Oct. 2011.

[3] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: A survey," *IEEE Wireless Commun.*, vol. 14, no. 2, pp. 70–87, Apr. 2007.

[4] G.-J. Qi, C. Aggarwal, D. Turaga, D. Sow, and P. Anno, "State-driven dynamic sensor selection and prediction with state-stacked sparseness," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Sydney, NSW, Australia, 2015, pp. 945–954.

[5] D. K. Jha, T. A. Wettergren, A. Ray, and K. Mukherjee, "Topology optimisation for energy management in underwater sensor networks," *Int. J. Control*, vol. 88, no. 9, pp. 1775–1788, Sep. 2015.

[6] Y. Li, D. K. Jha, A. Ray, and T. A. Wettergren, "Feature level sensor fusion for target detection in dynamic environments," in *Proc. Amer. Control Conf.*, Chicago, IL, USA, 2015, pp. 2433–2438.

[7] A. L. Buczak, H. H. Wang, H. Darabi, and M. A. Jafari, "Genetic algorithm convergence study for sensor network optimization," *Inf. Sci.*, vol. 133, nos. 3–4, pp. 267–282, 2001.

[8] F. Zhao, J. Shin, and J. Reich, "Information-driven dynamic sensor collaboration," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 61–72, Mar. 2002.

[9] X. Shen and P. Varshney, "Sensor selection based on generalized information gain for target tracking in large sensor networks," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 363–375, Jan. 2014.

[10] M. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks," *Int. J. High Perform. Comput. Appl.*, vol. 16, no. 3, pp. 293–313, 2002.

[11] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[12] E. F. Nakamura, A. A. F. Loureiro, and A. C. Frery, "Information fusion for wireless sensor networks: Methods, models, and classifications," *ACM Comput. Surveys*, vol. 39, no. 3, p. 9, 2007.

[13] H. B. Mitchell, *Multi-Sensor Data Fusion: An Introduction*. Heidelberg, Germany: Springer, 2007.

[14] P. Jesus, C. Baquero, and P. S. Almeidao, "A survey of distributed data aggregation algorithms," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 381–404, 1st Quart., 2015.

[15] N. Katenka, E. Levina, and G. Michailidis, "Local vote decision fusion for target detection in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 329–338, Jan. 2008.

[16] A. Arora *et al.*, "A line in the sand: A wireless sensor network for target detection, classification, and tracking," *Comput. Netw.*, vol. 46, no. 5, pp. 605–634, 2004.

[17] M. Zhu *et al.*, "Fusion of threshold rules for target detection in wireless sensor networks," *ACM Trans. Sensor Netw. (TOSN)*, vol. 6, no. 2, p. 18, 2010.

[18] Y. Li, D. K. Jha, A. Ray, and T. A. Wettergren, "Sensor selection for passive sensor networks in dynamic environment: A dynamic data-driven approach," in *Proc. Amer. Control Conf.*, Boston, MA, USA, Jul. 2016, pp. 4924–4929.

[19] C. Bishop, *Pattern Recognition*. New York, NY, USA: Springer, 2006.

[20] A. Ray, "Symbolic dynamic analysis of complex systems for anomaly detection," *Signal Process.*, vol. 84, no. 7, pp. 1115–1130, Jul. 2004.

[21] K. Mukherjee and A. Ray, "State splitting and merging in probabilistic finite state automata for signal representation and analysis," *Signal Process.*, vol. 104, pp. 105–119, Nov. 2014.

[22] S. Sarkar, D. K. Jha, A. Ray, and Y. Li, "Dynamic data-driven symbolic causal modeling for battery performance & health monitoring," in *Proc. 18th Int. Conf. Inf. Fusion (Fusion)*, Washington, DC, USA, 2015, pp. 1395–1402.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[24] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Phys. Rev. Lett.*, vol. 100, no. 15, 2008, Art. no. 158101.

[25] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, 2000.

[26] Y. Li, D. K. Jha, A. Ray, and T. A. Wettergren, "Information fusion of passive sensors for detection of moving targets in dynamic environments," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 93–104, Jan. 2017.

[27] D. K. Jha, A. Srivastav, K. Mukherjee, and A. Ray, "Depth estimation in Markov models of time-series data via spectral analysis," in *Proc. Amer. Control Conf. (ACC)*, Chicago, IL, USA, 2015, pp. 5812–5817.

[28] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[29] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, 1995.

[30] C. S. Daw, C. E. A. Finney, and E. Tracy, "A review of symbolic analysis of experimental data," *Rev. Sci. Inst.*, vol. 74, no. 2, pp. 915–930, 2003.

[31] Y. Seto, N. Takahashi, D. K. Jha, N. Virani, and A. Ray, "Data-driven robot gait modeling via symbolic time series analysis," in *Proc. Amer. Control Conf. (ACC)*, Boston, MA, USA, Jul. 2016, pp. 3904–3909.

[32] H. Liu, F. Hussain, C.-L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Min. Knowl. Disc.*, vol. 6, no. 4, pp. 393–423, 2002.

[33] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.

[34] U. M. Khepera, III, *Ver. 2.1*, K-Team SA, Vallorbe, Switzerland, 2008.

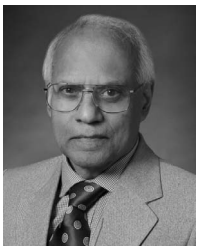
[35] H. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer-Verlag, 1994.

Yue Li received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2010, and the M.A. degree in mathematics and the Ph.D. degree in mechanical engineering from Pennsylvania State University, University Park, PA, USA, in 2016.

He is currently with FARO Technologies Inc., Orlando, FL, USA. His current research interests include system parameter identification and machine learning.

Devesh K. Jha (S'13–M'16) received the B.E. degree in mechanical engineering from Jadavpur University, Kolkata, India, in 2010, and the M.A. in mathematics and the Ph.D. degree in mechanical engineering from Pennsylvania State University, University Park, PA, USA, in 2016.

He is currently with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. His current research interests include machine learning, deep learning, and reinforcement learning.



Asok Ray (SM'83–F'02) received the M.S. degree in electrical engineering, mathematics, and computer science and the Ph.D. degree in mechanical engineering from Northeastern University, Boston, MA, USA.

He joined Pennsylvania State University, University Park, PA, USA, in 1985, where he is currently a Distinguished Professor of Mechanical Engineering and Mathematics, a Graduate Faculty of Electrical Engineering, and a Graduate Faculty of Nuclear Engineering. He has authored or co-authored over 600 research publications including about 300 scholarly articles in refereed journals and research monographs.

Dr. Ray is a fellow of ASME and World Innovative Foundation.

Thomas A. Wettergren (A'95–M'98–SM'06) received the B.S. degree in electrical engineering and the Ph.D. degree in applied mathematics from Rensselaer Polytechnic Institute, Troy, NY, USA.

He joined the Naval Undersea Warfare Center, Newport, RI, USA, in 1995, where he has served as a Research Scientist in the torpedo systems, sonar systems, and undersea combat systems departments. He currently serves as the U.S. Navy Senior Technologist for Operational and Information Science as well as a Senior Research Scientist with the Newport Laboratory.

Dr. Wettergren is a member of the Society for Applied and Industrial Mathematics.