This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Efficient Sensor Selection for Active Information Fusion

Yongmian Zhang, *Member, IEEE*, and Qiang Ji, *Senior Member, IEEE*

***Abstract*—In our previous paper, we formalized an active information fusion framework based on dynamic Bayesian networks to provide active information fusion. This paper focuses on a central issue of active information fusion, i.e., the efficient identification of a subset of sensors that are most decision relevant and cost effective. Determining the most informative and cost-effective sensors requires an evaluation of all the possible subsets of sensors, which is computationally intractable, particularly when information-theoretic criterion such as mutual information is used. To overcome this challenge, we propose a new quantitative measure for sensor synergy based on which a sensor synergy graph is constructed. Using the sensor synergy graph, we first introduce an alternative measure to multisensor mutual information for characterizing the sensor information gain. We then propose an approximated nonmyopic sensor selection method that can efficiently and near-optimally select a subset of sensors for active fusion. The simulation study demonstrates both the performance and the efficiency of the proposed sensor selection method.**

***Index Terms*—Active information fusion, Bayesian networks (BNs), sensor selection, situation awareness.**

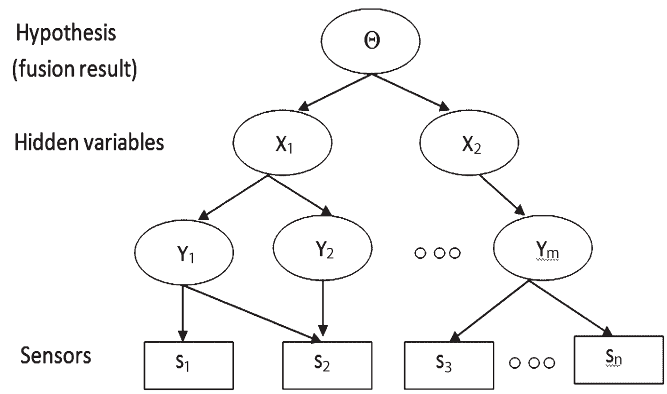

Fig. 1. BN is used for active information fusion, where Θ and S_i are hypothesis and sensors, respectively. X_i and Y_i are the intermediate variables, and they are needed to model the relationships among sensors and the hypothesis. Sensor fusion is accomplished through probabilistic inference given the sensory measurements.

I. Introduction

INFORMATION fusion is playing an increasingly important role in improving the performance of sensory systems for various applications, including situation assessment, enemy intent understanding and prediction, and threat assessment. As sensors become ubiquitous, persistent, and pervasive, and coupled with the ever increasing demand for less time and fewer resources, it becomes critically important to perform selective fusion so that decision can be made in a timely and efficient manner. The need for sensor selection is further demonstrated by the availability of an increasingly large volume of sensory data and by the variability of sensor reliability over time and over location. It is important to select the sensors not only to reduce the amount of data to integrate but also to improve fusion accuracy by selecting the most reliable sensors for a certain location at a certain time, by selecting complementary sensors, and by reducing sensor redundancy. Active fusion serves these purposes well. Active fusion extends the paradigm of information fusion by being not only concerned with the methodology of how to combine information but also concerned with the fusion efficiency, timeliness, and accuracy. Active fusion can be defined as the process of combining data with a control mechanism that dynamically selects a subset of sensors to minimize uncertainty in situation assessment and to maximize the overall expected utility in decision making.

In our previous work [1], we formalized an information fusion framework based on Bayesian networks (BNs) to provide active and sufficing information fusion. BNs are used to model a number of uncertain events, their spatial relationships, and the sensor measurements. Given the sensory measurements, information fusion is performed through probabilistic inference using the BN. This can be accomplished through bottom-up belief propagation, as illustrated in Fig. 1. Our previous work, however, did not address the core issue in active fusion, i.e., efficient sensor selection. This is the focus of this paper.

Based on information theory [2], the more sensors[1] we use, the more information we can obtain. However, every act of information gathering incurs cost. Sensor costs may include physical costs, computational costs, maintenance costs, and human costs (e.g., risk). Many applications are often constrained by limited time and resources. An essential issue for active information fusion is to select a subset of the most

Manuscript received May 20, 2008; revised October 25, 2008, December 19, 2008, and March 11, 2009. This work was supported in part by the Air Force Office of Scientific Research under Grant F49620-03-0160. This paper was recommended by Associate Editor R. Lynch.

The authors are with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: zhangy@ecse.rpi.edu; qji@ecse.rpi.edu).

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/TSMCB.2009.2021272

[1]For generality, sensors could refer to any devices/means of acquiring information. For example, they may be electromagnetic or acoustic devices or they could also be direct observations of the world through reconnaissance and intelligence gathering activities.

synergetic sensors, which can maximally reduce the uncertainty about the events of interest with minimum costs. Dynamically determining the best set of sensors, given the uncertainty about the state of the world, requires to enumerate all the possible subsets of sensors, which is computationally intractable and practically infeasible. This computational difficulty is twofold. First, the computation of a sensor selection criterion such as mutual information is exponential with respect to the number of sensors. Second, searching for an optimal subset of sensors is also NP-hard, since the sensor space exponentially increases with the number of sensors. To address this computational difficulty, a common practice is to use myopic analysis, which assumes that only one observation will be available at a time, even when there is an opportunity to make a set of observations [3]–[6]. There is a vast literature on the problem of single optimal sensor selection [7]–[9]. However, the myopic approach cannot guarantee to obtain the best evidences that most effectively reduce uncertainty and cost. To effectively reduce uncertainty and cost, one should use nonmyopic selection, which simultaneously considers several observations before making a decision. The most common nonmyopic method is the greedy approach. While efficient, it cannot guarantee optimality with the selected sensors. Other works try to overcome the limitations with the greedy approach, yet with their own strong assumptions. In [10], Heckerman *et al.* presented an approximate nonmyopic approach based on the central-limit theorem in an influence diagram (ID) for efficiently computing the value of information. Their method, however, assumes that the sensors are conditionally independent of each other, given the decision variable, and that the decision variable is binary. Krause and Guestrin [11], [12] presented a randomized approximation algorithm for selecting a near-optimal subset of observations for graphical models. Under the assumptions that the sensors are conditionally independent given the decision variables, the information gain is then guaranteed to be a submodular function, and the theory of submodular functions can then be applied to achieve a near-optimal solution in selecting a subset of observations using a greedy approach. Recently, Liao and Ji [13] have presented an approximation algorithm for the nonmyopic computation of the value of information in an ID. Their method extends the approach in [10] without requiring the sensors being conditionally independent of each other and the decision node being binary.

This paper takes another avenue of approach to efficiently select a subset of near-optimal sensors without the strong sensor independence assumptions, as made in [10] and [12]. Specifically, we first introduce a new quantitative measure of sensor synergy based on mutual information. Based on the synergy measure, we then introduce a method to efficiently compute the least upper bound (LUB) of mutual information for a set of sensors. Experiments show that the LUB closely approximates the mutual information in value, as shown in Figs. 5 and 6. Hence, the computational difficulty with computing the exact nonmyopic mutual information can, therefore, be circumvented by computing its LUB instead. In addition, the synergy measure can also be used to prune the sensor space, which, therefore, reduces the search time for the best sensor set. A summary of the mentioned work may be found in [14].

II. Problem Formulation

The problem of sensor selection for active fusion can be stated as follows: Assume that there are m sensors S_i, $i = 1,\ldots,m$, available that provide measurements of the world. Let Θ be a set of hypothesis θ_k of the world situation $k = 1,\ldots,K$. Let $\mathbf{S} = \{S_1,\ldots,S_n\}$ be a subset of n sensors selected at time t, where $n \in \{1,\ldots,m\}$. Let $C(\mathbf{S})$ be the cost to use the set of sensors $\mathbf{S}$. The objective of sensor selection at time $t+1$ is to select a subset of sensor $\mathbf{S}^*$ to achieve the maximal utility, i.e.,

$$\mathbf{S}^* = \arg\max_{\mathbf{S}\in\mathcal{S}} U(u_1,u_2) \tag{1}$$

where u_1 and u_2 denote information gain (i.e., the mutual information) and the sensor usage cost saving, respectively, $\mathcal{S}$ represents all the possible subsets of sensors, and $U(u_1,u_2)$ is a utility function. Here, we use $u_2 = 1 - C(\mathbf{S})$ to convert the sensor usage cost to the corresponding cost saving, which makes u_1 and u_2 qualitatively equivalent. For simplicity, in this paper, we assume that the cost is the same for all sensors. Hence, we can ignore u_2.

The major difficulty of using (1) for sensor selection is to efficiently compute the information gain u_1. From information theory, the entropy of hypothesis Θ given a sensor S_i measures how much uncertainty exists in Θ given S_i, i.e.,

$$H(\Theta\mid S_i) = -\sum_{s_i}\sum_{\Theta} P(\theta,s_i)\log P(\theta\mid s_i) \tag{2}$$

where s_i denotes a reading of sensor S_i.[2] Subtracting $H(\Theta\mid S_i)$ from the original uncertainty in Θ without S_i, i.e., $H(\Theta)$, yields the expected amount of information about Θ that S_i is capable of providing

$$\begin{aligned} I(\Theta;S_i) &= H(\Theta) - H(\Theta\mid S_i)\\ &= -\sum_{\Theta} P(\theta)\log P(\theta)\\ &\quad + \sum_{S_i}\left\{P(s_i)\sum_{\Theta} P(\theta\mid s_i)\log P(\theta\mid s_i)\right\}\\ &= \sum_{\Theta}\sum_{S_i} P(\theta,s_i)\log\frac{P(\theta\mid s_i)}{P(\theta)} \end{aligned} \tag{3}$$

where $I(\Theta;S_i)$ is referred to as the mutual information, which characterizes the expected total uncertainty-reducing potential of Θ due to S_i. The mutual information for a sensor set $\mathbf{S} = \{S_1,\ldots,S_n\}$ can be obtained by

$$\begin{aligned} &I(\Theta;\mathbf{S})\\ &= H(\Theta) - H(\Theta\mid\mathbf{S})\\ &= -\sum_{\Theta} P(\theta)\log P(\theta)\\ &\quad + \sum_{\Theta}\sum_{S_1}\cdots\sum_{S_n}\{P(\theta,s_1,\ldots,s_n)\log P(\theta\mid s_1,\ldots,s_n)\}\\ &= \sum_{\Theta}\sum_{S_1}\cdots\sum_{S_n}\left\{P(\theta,s_1,\ldots,s_n)\log\frac{P(\theta\mid s_1,\ldots,s_n)}{P(\theta)}\right\} \end{aligned} \tag{4}$$

[2]Without loss of generality, here we assume discrete sensor measurement. The theories can be straightforwardly extended to continuous sensor measurements.

where $P(\theta, s_1, \ldots, s_n)$ and $P(\theta \,|\, s_1, \ldots, s_n)$ at time t can directly be obtained through BN inference. The mutual information in (4) provides a sensor selection criterion in terms of the uncertainty reduction potential, i.e., mutual information.

It is clear from (4) that when the number of sensors in $\mathbf{S}$ is large or when the number of states for each sensor is large, it becomes computationally impractical to simply implement this information-theoretic criterion, because it generally requires time exponential in the number of summations to exactly compute the mutual information. The remainder of this paper addresses this computational difficulty.

III. Approximation Algorithm

In this section, we give a graph-theoretic definition of sensor synergy. We then present the theorems on which our algorithm is based.

A. Sensor Synergy in Information Gain

Throughout this section, it is assumed that we have obtained $I(\Theta; S_i, S_j)$ and $I(\Theta; S_i)$, i.e., the mutual information of all pairs of sensors and individual sensors with respect to Θ, respectively. We will introduce an efficient method to obtain all $I(\Theta; S_i, S_j)$ in Section III-C. We first define a synergy coefficient to characterize the synergy between two sensors, and then extend this definition to multiple sensors.

Definition 1 (Synergy Coefficient): A measure of the expected synergetic potential between two sensors S_i and S_j in reducing the uncertainty of hypothesis Θ is defined as

$$r_{ij} = \frac{I(\Theta; S_i, S_j) - \max(I(\Theta; S_i), I(\Theta; S_j))}{H(\Theta)}. \quad (5)$$

The denominator $H(\Theta)$ in (5) is to restrict r_{ij} to the interval [0, 1]. It can easily be proved that $r_{ij} \geq 0$ based on the "information never hurts" principle [2], i.e., $I(\Theta; S_i, S_j) \geq I(\Theta; S_i)$, and $I(\Theta; S_i, S_j) \geq I(\Theta; S_j)$. This follows that S_i and S_j taken together are always more informative than when they are taken alone. The larger r_{ij} is, the more synergetic the sensors S_i and S_j are. Obviously, $r(\cdot, \cdot)$ is symmetrical in S_i and S_j, and $r_{ij} = 0$ if $i = j$.

Definition 2 (Synergy Matrix): Let a sensor set be $\mathbf{S} = \{S_1, \ldots, S_n\}$. The sensor synergy matrix is an $n \times n$ matrix defined as

$$R = \begin{bmatrix} 0 & r_{12} & \cdots & r_{1n} \\ r_{21} & 0 & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & 0 \end{bmatrix}. \quad (6)$$

R is an information measure of synergy among sensors that is based on pairwise sensor synergy. With a synergy matrix, we naturally define its graphical representation.

Definition 3 (Synergy Graph): Given a sensor synergy matrix, a graph $G = (\mathbf{S}, \mathbf{E})$, where $\mathbf{S}$'s are the nodes representing the set of available sensors, and $\mathbf{E}$'s are the links representing the set of pairwise synergetic links weighted by synergy coefficients r_{ij}, is a sensor synergy graph.

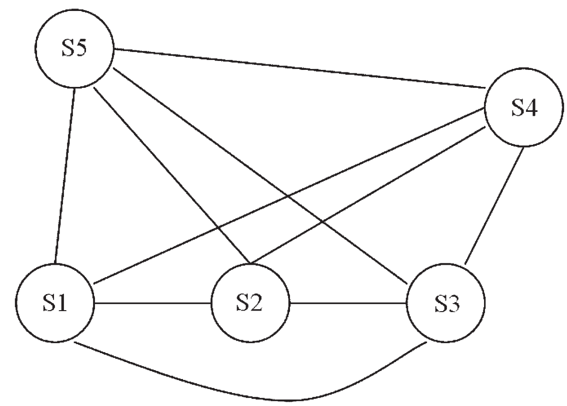

Fig. 2. Example of synergy graph with five sensors.

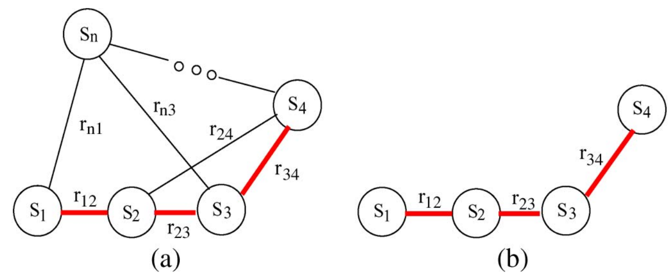

Fig. 3. (a) Synergy chain $\{S_1, S_2, S_3, S_4\}$ (highlighted) on a pruned synergy graph. (b) Corresponding MSC.

We use the synergy graph to graphically represent the synergy among multiple sensors. By definition of the synergy, G is a complete graph, i.e., there is a link between any two nodes in the graph. Fig. 2 gives an example synergy graph consisting of five sensors.

Definition 4 (Pruned Synergy Graph): A pruned synergy graph is created from a synergy graph after removing some links. A pruned synergy is, therefore, not a complete graph.

Fig. 3 shows an example of a pruned synergy graph. To further exploit the theoretical properties of mutual information $I(\Theta; \mathbf{S})$ for a set of sensors, we give the following definitions.

Definition 5 (Synergy Chain): Given a pruned synergy graph G, if all the sensors in a subset on G are serially linked, then this subset of sensors is referred to as a sensor synergy chain. Note that while the sensors in a set $\mathbf{S}$ are generally order independent, the sensors in a synergy chain are order dependent and sequentially ordered.

Definition 6 (MSC): Given a synergy chain with n sensors, for all $i = 1, \ldots, n-1$, if $p(S_{i+1} \,|\, S_1, S_2, \ldots, S_i) = p(S_{i+1} \,|\, S_i)$, then the chain that describes the synergetic relationship among $\{S_1, \ldots, S_n\}$ is a Markov synergy chain (MSC). An MSC is also ordered.

Fig. 3 graphically shows the above definitions about the synergy chain in a pruned synergy graph. The MSC represents an ideal synergetic relationship among sensors. The MSC rarely exists in practice, but this does not prevent us from using it as a basis for the graph-theoretic analysis of synergy among sensors. In fact, as to be shown later, the concept of MSC is used to define the upper bound for the mutual information of a set of sensors. With the above definitions, we give the following two theorems.

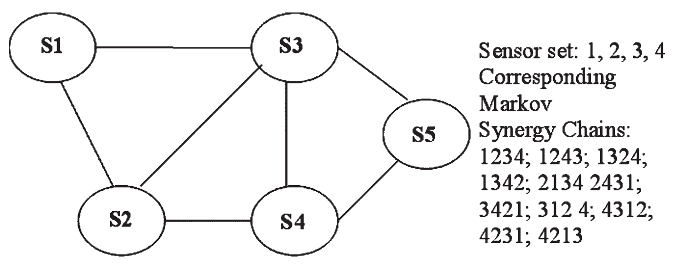

Fig. 4. Illustration of a set of possible MSCs for a set unordered four sensors in a pruned synergy graph.

Theorem 1 (MSC Rule): Given an MSC with a set of ordered sensors $\mathbf{S} = \{S_1, \ldots, S_n\}$, for any n, the joint mutual information with respect to Θ for sensors on an MSC is

$$I^M(\Theta; S_1, \ldots, S_n) = I(\Theta; S_1) + \sum_{i=1}^{n-1} \left(I(\Theta; S_i, S_{i+1}) - I(\Theta; S_i)\right). \quad (7)$$

The proof of this theorem can be found in Appendix A. We want to make note that the mutual information for an MSC is sensor-order dependent due to the pairwise synergy definition. The significance of Theorem 1 is that it allows us to efficiently compute the joint mutual information for n ($n > 2$) ordered sensors as a sum of mutual information of only singleton and pairwise sensors if the set of sensors forms an MSC. In contrast to (4), the computational cost of (7) is dramatically reduced. Although (7) is particularly for an MSC, the theorem above has some useful properties that can be used for the solution of our sensor selection problem.

Theorem 2 (Synergy Upper Bound): For a set of unordered sensors $\mathbf{S} = \{S_1, \ldots, S_n\}$, its mutual information is upper bounded by the mutual information of the corresponding MSC, i.e.,

$$I(\Theta; S_1, \ldots, S_n) \leq I^M(\Theta; S_1, \ldots, S_n). \quad (8)$$

The proof of this theorem is provided in Appendix B. Please note that while $I(\Theta; S_1, \ldots, S_n)$ is sensor-order independent, $I^M(\Theta; S_1, \ldots, S_n)$ is sensor-order dependent. As a result, depending on the order of sensors in $\mathbf{S}$, different MSCs may be produced. Let

$$I^M_{\min} = \arg\min_{\mathcal{S}} \left(I^M(\Theta; \mathcal{S})\right)$$
$$I^M_{\max} = \arg\max_{\mathcal{S}} \left(I^M(\Theta; \mathcal{S})\right) \quad (9)$$

where $\mathcal{S}$ denotes all the possible orders of a sensor set $\{S_1, \ldots, S_n\}$. $I^M_{\min}$ is referred to as the LUB of $I(\Theta; S_1, \ldots, S_n)$, and $I^M_{\max}$ is referred to as the greatest upper bound (GUB) of $I(\Theta; S_1, \ldots, S_n)$. For example, in Fig. 4, the sensor set $\mathbf{S} = \{S_1, S_2, S_3, S_4\}$ has multiple MSCs, as given in this figure, and there exist a LUB and a GUB of $I(\Theta; \mathbf{S})$.

We are particularly interested in the LUB of $I(\Theta; \mathbf{S})$ due to two reasons. First, it can be seen from Figs. 5 and 6 that the LUBs of $I(\Theta; \mathbf{S})$ closely follow the trend of $I(\Theta; \mathbf{S})$ in the entire space of sensor subsets. Second, the exact value of $I(\Theta; \mathbf{S})$ and its LUB are quantitatively very close in value. Thus, $I^M_{\min}(\Theta; \mathbf{S})$ provides a substitute measure for $I(\Theta; \mathbf{S})$ that can be used to evaluate an optimal sensor subset. Importantly, the LUBs of $I(\Theta; \mathbf{S})$ can simply be written as the sum of the mutual information of only pairwise sensors and singleton sensors, as shown in (7), hence, with relatively very low computational cost. Therefore, the computational difficulty in exactly computing the higher-order mutual information can be circumvented by only computing the LUBs of the mutual information. This is the central strategy of our approach.

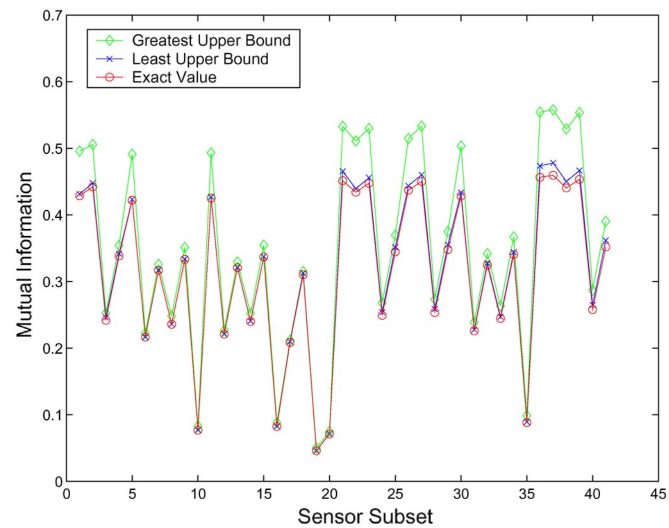

Fig. 5. Bound of mutual information $I(\Theta, S)$ and its exact value from a six-sensor BN model. The X-axis represents the indexes of 41 sensor subsets. Labels 1–20 are the indexes of the three-sensor subsets; Labels 21–35 are the indexes of the four-sensor subsets; and Labels 36–41 are the indexes of the five-sensor subsets.

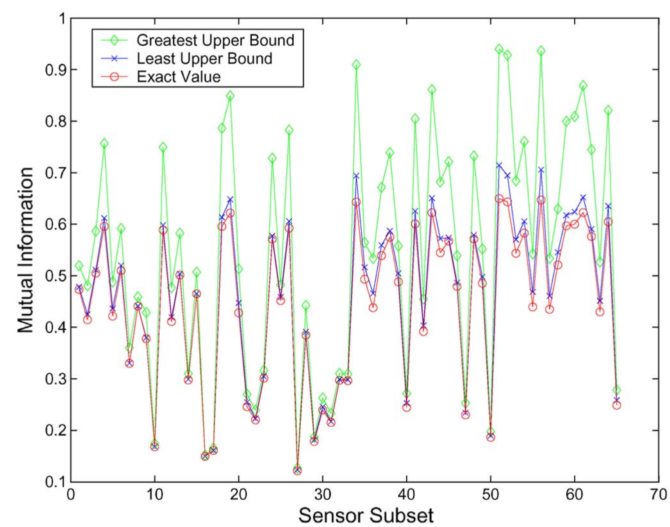

Fig. 6. Bound of mutual information $I(\Theta, S)$ and its exact value from a ten-sensor BN model. The X-axis represents the indexes of sensor subsets. For clarity, the figure only shows 66 subsets out of 627. Labels 1–18 are the indexes of the five-sensor subsets; Labels 19–34 are the indexes of the six-sensor subsets; Labels 35–51 are the indexes of the seven-sensor subsets; and Labels 52–66 are the indexes of the eight-sensor subsets.

B. Pruning Synergy Graph

The synergy graph is a completely connected network due to the weights of synergy graph $r_{ij} \geq 0$. Some sensors are highly synergetic, whereas others are not. Intuitively, sensors that

TABLE I
EXAMPLE OF SYNERGY COEFFICIENT WITHOUT PRUNING

r_{ij}	1	2	3	4	5	6	7	8	9	10
1	0	0.0004	0.0034	0.0034	0.0029	0.0023	0.0035	0.0035	0.0031	0.0017
2		0	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0002
3			0	0.0215	0.0196	0.0156	0.0099	0.0122	0.0212	0.0113
4				0	0.0184	0.0147	0.0099	0.0122	0.0198	0.0105
5					0	0.0930	0.0083	0.0104	0.0650	0.0659
6						0	0.0066	0.0082	0.0517	0.1329
7							0	0.0100	0.0091	0.0050
8								0	0.0112	0.0060
9									0	0.0388
10										0

cause a very small reduction in uncertainty of hypotheses are those that give us the least additional information beyond what we would obtain from other sensors. In such cases, r_{ij} is very small. We prune the sensor synergy graph so that many weak sensor combinations are eliminated while preserving the most promising ones. This can significantly reduce the search space in identifying the optimal sensor subset. We prune the synergy matrix (the weights of the synergy graph) in (6) by using

$$r_{ij} = \begin{cases} 1, & r_{ij} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where τ is a pruning threshold. The selection of an appropriate threshold τ is problem dependent. We want to note that although there is no theoretical basis to determine a good pruning threshold, our empirical tests, however, show that using the arithmetic average of r_{ij} as the pruning threshold can preserve most of the strong synergetic connections in the graph while eliminating weak links. After pruning, a fully connected synergy graph then becomes a sparse graph. Table I and

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ & & & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ & & & & 0 & 1 & 0 & 0 & 1 & 1 \\ & & & & & 0 & 0 & 0 & 1 & 1 \\ & & & & & & 0 & 0 & 0 & 0 \\ & & & & & & & 0 & 0 & 0 \\ & & & & & & & & 0 & 1 \\ & & & & & & & & & 0 \end{bmatrix} \tag{11}$$

are examples of the synergy coefficient before and after pruning. Fig. 7 illustrates their corresponding synergy graph from a completely connected network to a sparse graph after pruning.

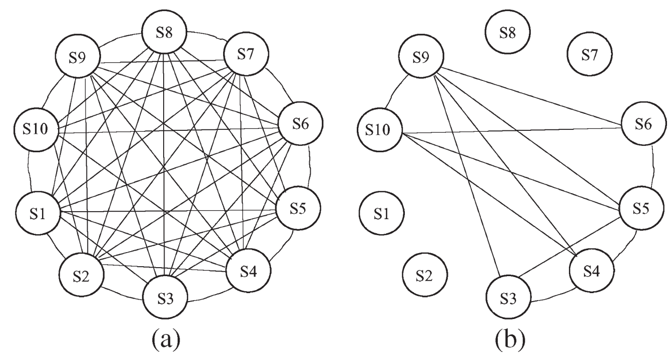

Fig. 7. (a) Completely connected synergy graph and the links are weighted by r_{ij}, as shown in Table I. (b) Pruned synergy graph and its corresponding matrix as shown in (11). The pruning threshold is the average of r_{ij}, and it is 0.0161.

C. Computing Pairwise Mutual Information

In the above sections, we assumed that we have known the mutual information of the pairwise sensors $I(\Theta; S_i, S_j)$. For n sensors, there are $(n(n-1)/2)$ pairs of sensors. To obtain the mutual information for one pair of sensors, it requires four repetitions of inferences if the sensor state is binary. Therefore, $2n(n-1)$ repetitions of inference are needed for all pairs of sensors. Although this computation is manageable, it still severely limits the performance as n becomes large. Fortunately, there is an efficient way to compute the mutual information for all pairs of sensors [15], [16].

Referring to Fig. 1, the joint probability of hypothesis Θ and pairwise sensors $\{S_i, S_j\}$ may be written as in (12), shown at the bottom of the next page, where $\pi(x)$ represents the parental nodes of node x. From (12), it can be observed that the first factor $P(\Theta)\prod_{k=1}^{K} P(X_k \mid \pi(X_k)) \prod_{m=1}^{M} P(Y_m \mid \pi(Y_m))$ is related to the part of the BN structure that does not include the sensors. The structure is, therefore, fixed, and so are its probabilities. Hence, this term is constant, independent of the pair of sensors used. On the other hand, the second factor $\{\sum_{S_1, S_l, \ldots, S_N, l\neq i\ l\neq j} \prod_{n=1}^{N} P(S_n \mid \pi(S_n))\}$ varies, depending on the pair of sensors selected. Therefore, we do not need to recalculate the unchanged part (the first factor) of (12) at each time. Instead, we only need to compute it once for all pairs of sensors, but use it over time so that the computation of pairwise mutual information can significantly be curtailed. Given $P(\Theta, S_i, S_j)$, it can then be substituted into (4) to compute $I(\Theta; S_i, S_j)$. Details of this method can be found in [15] and [16]. Fig. 8 illustrates the comparative result of time saving in computing (12) for all pairs of sensors by using our method and by directly using two inference algorithms, namely, clique tree propagation (CTP) [17], [18] and variable elimination (VE) [19]. The evaluation is performed on a six-layer BN model with 10, 15, and 20 sensors.

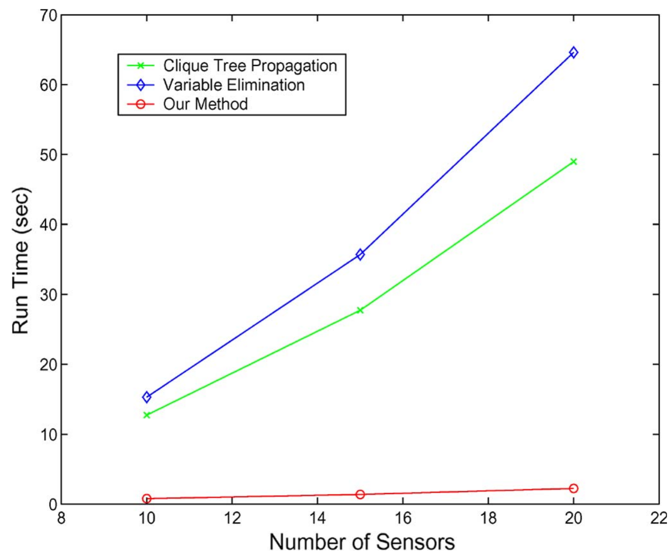

Fig. 8. Comparison of time saving among CTP, VE, and our method in computing (12) for all pairs of sensors. It can be seen that our method can significantly save time.

D. Approximation Algorithm

We are now ready to provide the complete algorithm. Let $\mathbf{S}$ denote the current set of selected sensors, and let $lub(\Theta; \mathbf{S})$ be the LUB of $I(\Theta; \mathbf{S})$. The approximation sensor selection algorithm is given in Table II. Guided by the pruned synergy graph, the algorithm starts with the best pair of sensors identified through an exhaustive search and then searches for the next best sensor. The next best sensor is the one, when added to the current sensor ensemble, that yields the highest utility, which is computed from $lub(\Theta; \mathbf{S})$. This process repeats, with one sensor added to the current sensor ensemble at a time, until the newly added sensor does not yield an improvement in sensor utility. Although the algorithm is greedy, the searching process is guided by a synergy graph so that the selected sensor subset is serially connected. This, therefore, ensures both the quality and the speed of sensor selection.

TABLE II
PSEUDOCODE OF THE APPROXIMATION ALGORITHM TO SELECT A SUBSET OF SENSORS

```
SENSOR-SELECTION(n)
for each i, j, compute I(Θ; S_i) and I(Θ; S_i, S_j)
Construct a pruned synergy graph G
Choose S_{i*}, S_{j*} such that I(Θ; S_i, S_j)
    is maximized for all i and j
S ← {S_{i*}, S_{j*}}
while |S| < m
    for each S', where |S'| = |S| + 1, and S' is a synergy
        chain on G and S ⊂ S'
        Find all Markov synergy chains of S'
        lub(Θ; S') ← arg min I(Θ; S'),
            where I(Θ; S') is computed by Eq. (7), and
            min is taken over all Markov synergy chains of S'
    S'* ← arg max lub(Θ; S')
        where max is taken over all S'
    if lub(Θ; S'*) > lub(Θ; S)
        S ← S'*
    else break
return S
```

IV. ALGORITHM EVALUATION

Since the main contribution of this paper is the introduction of an alternative measure to mutual information for efficient sensor selection, the experimental evaluation should focus on the effectiveness of this measure for both sensor selection accuracy and efficiency. We want to emphasize that the alternative measure, i.e., the LUB of mutual information, is an approximation of the mutual information only for the purpose of sensor selection. As a result, the quality of this approximation should be evaluated against its performance in sensor selection. For this, we propose to measure how close the sensor selection results using the alternative measure are to those based on mutual information. The closeness between a sensor subset selected using the alternative measure and a sensor subset selected based on mutual information is quantified by the relative difference in mutual information. Based on this criterion, we will experimentally evaluate the proposed method under different BN topologies, different BN model complexities, and different number of sensors.

Given two different criteria (mutual information and its LUB) for measuring sensor gain, sensor selection can be carried out by using different methods. We will perform sensor selection using the following methods: 1) brute-force method; 2) random method, which randomly chooses one sensor at a time to form a sensor ensemble; and 3) the proposed method. These experiments try to demonstrate the following: 1) The proposed LUB criterion suboptimally works for different methods. 2) Given the same sensor selection criterion, the proposed greedy approach outperforms the random sensor selection method.

$$
\begin{aligned}
P(\Theta, S_i, S_j) &= \sum_{S_1, S_l, \ldots, S_n, l \neq i\ l \neq j} \left\{ P(\Theta) \prod_{k=1}^{K} P\left(X_k \mid \pi(X_k)\right) \prod_{m=1}^{M} P\left(Y_m \mid \pi(Y_m)\right) \prod_{n=1}^{N} P\left(S_n \mid \pi(S_n)\right) \right\} \\
&= P(\Theta) \prod_{k=1}^{K} P\left(X_k \mid \pi(X_k)\right) \prod_{m=1}^{M} P\left(Y_m \mid \pi(Y_m)\right) \left\{ \sum_{S_1, S_l, \ldots, S_n, l \neq i\ l \neq j} \prod_{n=1}^{N} P\left(S_n \mid \pi(S_n)\right) \right\}
\end{aligned}
\tag{12}
$$

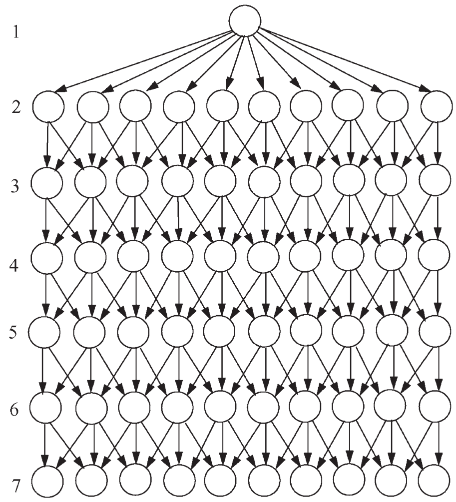

Fig. 9. Generic example of the BN network used for evaluation, where the top layer is for the hypothesis, and the bottom layer is for the sensors. The intermediate layers are arbitrarily and randomly connected.

We first compare the performance of the proposed sensor selection method in Table II with the brute-force method. The brute-force method exhaustively identifies the best sensor subset by the exact mutual information. The study is done by using different numbers of sensors and different BN topologies. Fig. 9 shows a generic example of a BN used for the evaluation.

Due to the exponential time with the brute-force approach, we limit our test models to up to five layers and up to ten sensors or less. The exact number of layers, the connections among nodes in the intermediate layers, and the number of sensors are randomly generated so that ten different BNs with different topologies are generated. For each randomly generated BN topology, its parameters are randomly parameterized ten times to produce ten differently parameterized BNs for each selected topology. This yields a total of 100 test models. Fig. 10 shows two examples of BNs used for this paper.

The results averaged among 100 trials are shown in Table III, where the closeness is defined as the relative difference in mutual information between the solution from our approach and the solution from the brute-force approach. It can be seen that the solution with our method is close to the sensor selection results using the brute-force method. For further comparison, the run time of the two methods is measured on a 2.0-GHz computer, and the run time averaged among ten trials is summarized in Table III. Compared with the brute-force method, our method significantly reduces the computation time with minimum loss in sensor selection accuracy.

To demonstrate the improvement of the proposed method over random sensor selection, the results of random sensor selection are also included in Table III. For a fair comparison, we first use our method to select a best subset and then use the random method to select a subset of the same size using the same criterion. To account for the random nature of random selection, the results are averaged, and the averaged result is used to compare against the result from our method. Compared with the random sensor selection, our method shows a significant improvement in sensor selection accuracy.

Finally, we want to note that the randomly generated BN topologies (for example, the BNs in Figs. 9 and 10) may not necessarily satisfy the assumption needed for Theorem 1 to hold. Despite this, the selected sensors remain close (in mutual information) to those selected by the brute-force method, as demonstrated in Table III. We also repeat the above experiments by using naive BNs. The parameters of BNs are randomly generated. We selected k sensors from n sensors $(k < n)$ without considering sensor costs. The sensors selected by the brute-force method and by our approach have no difference.

V. Conclusion

It is computationally difficult to identify an optimal sensor subset with the information-theoretic criterion. To address problem, we have presented an approximation method to find a near-optimal sensor subset by utilizing the sensor pairwise information to infer the synergy among sensors. Specifically, this paper includes the following aspects: First, we propose to use a BN to represent sensors, their dependencies, and their relationships to other latent variables. In addition, the built-in conditional independence assumptions with the BNs allow factorizing the joint probabilities so that fusion can efficiently be performed. Second, we introduce a statistical measure to quantify the pairwise synergy among sensors. Based on the synergy measure, a synergy graph is constructed, which is used to infer synergy among multiple sensors, based on which we can then eliminate many unpromising sensor combinations. Finally, for the remaining sensor combinations, a greedy approach is introduced to identify the optimal sensor combination based on the LUB of the joint mutual information. The use of the LUB of the joint mutual information instead of the joint mutual information itself significantly reduces the computation time with minimum loss in accuracy. We demonstrate both the optimality and the efficiency of the proposed method through many random simulations under different numbers of sensors and different relationships among sensors.

A major assumption of this paper is that the two sensors are conditionally independent of each other, given another sensor between the two sensors and the fusion result. This assumption could limit the utility of this paper. As part of the future research, we will study ways to overcome this assumption. Another assumption we made in this paper is that all the sensors have the same cost. Such an assumption is not realistic for many applications. Overcoming this assumption, however, requires incorporating the sensor cost into the proposed synergy function, which is a nontrivial task. We will study this issue in the future as well.

Appendix

In the following, we introduce our proof for Theorems 1 and 2.

A. *Proof of Theorem 1*

Before proving Theorem 1, we give the following lemma.

Lemma 1 (Chain Rule of Mutual Information): Letting X, $Y_1, \ldots, Y_m$ be random variables, then

$$I(X; Y_1, \ldots, Y_m) = I(X; Y_1) + \sum_{i=2}^{M} I(X; Y_i \,|\, Y_1, \ldots, Y_{i-1}). \tag{13}$$

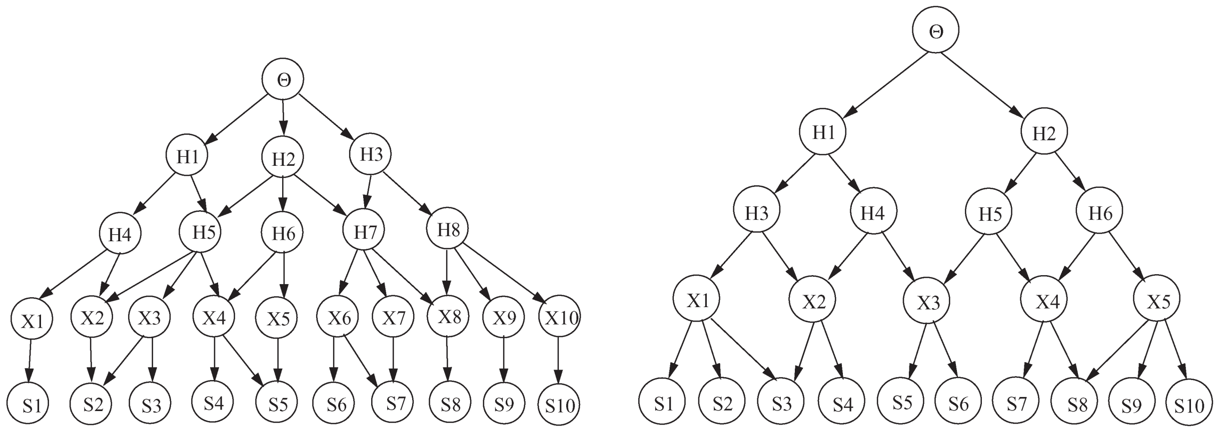

Fig. 10. Two specific examples of BN structures with different numbers of sensors used for the evaluation.

TABLE III
COMPARISON OF THE PROPOSED METHOD AND THE BRUTE-FORCE METHOD

Number of Sensors	Our Approach		Random Method	Brute-Force
	Relative mutual information difference of our method to brute force methods	Run time (Seconds)	Relative mutual information difference of random method to brute force methods	Run time (Seconds)
7	1.56%	1.020	21.13%	63.87
8	1.77%	1.099	28.32%	355.05
9	2.75%	1.209	36.54%	2967.36
10	1.89%	1.430	39.19%	13560.54

The proof of Lemma 1 is straightforward [2]. We now turn to proving Theorem 1.

Proof: Based on Lemma 1, we have

$$\begin{aligned} &I(\Theta; S_1, \ldots, S_m)\\ &\quad = I(\Theta; S_1) + I(\Theta; S_2 \mid S_1) + I(\Theta; S_3 \mid S_1, S_2)\\ &\qquad + I(\Theta; S_4 \mid S_1, S_2, S_3) + \cdots + I(\Theta; S_m \mid S_1, \ldots, S_{m-1}). \end{aligned} \tag{14}$$

We start with an MSC containing four random variables $\{\Theta, S_1, S_2, S_3\}$, then extend it to five variables, and finally to a finite number of arbitrary random variables forming an MSC. Notice that Θ is the hypothesis, and S_1, S_2, S_3 are the sensors.

Based on Definition 6 of MSC, S_1 and S_3 are conditionally independent given S_2, i.e., $P(S_3 \mid S_1, S_2) = P(S_3 \mid S_2)$. First, we prove that the following equation holds when S_1 and S_3 are conditionally independent given S_2:

$$I(\Theta; S_3 \mid S_2) = I(\Theta; S_3 \mid S_1, S_2) \tag{15}$$

$$\begin{aligned} &I(\Theta; S_3 \mid S_1, S_2)\\ &\quad = \sum_{\Theta, S_1, S_2, S_3} p(\theta, s_1, s_2, s_3) \left\{ \lg \frac{p(\theta, s_3 \mid s_1, s_2)}{p(\theta \mid s_1, s_2) p(s_3 \mid s_1, s_2)} \right\}\\ &\quad = \sum_{\Theta, S_1, S_2, S_3} p(\theta, s_1, s_2, s_3) \left\{ \lg \frac{p(\theta, s_3, s_1, s_2)}{p(\theta, s_1, s_2) p(s_3 \mid s_2)} \right\}\\ &\quad = \sum_{\Theta, S_1, S_2, S_3} p(\theta, s_1, s_2, s_3) \left\{ \lg \frac{p(s_3 \mid \theta, s_1, s_2)}{p(s_3 \mid s_2)} \right\}\\ &\quad = \sum_{\Theta, S_1, S_2, S_3} p(\theta, s_1, s_2, s_3) \left\{ \lg \frac{p(s_3 \mid \theta, s_2)}{p(s_3 \mid s_2)} \right\}\\ &\quad = \sum_{\Theta, S_2, S_3} p(\theta, s_2, s_3) \left\{ \lg \frac{p(s_3 \mid \theta, s_2)}{p(s_3 \mid s_2)} \right\}\\ &\quad = \sum_{\Theta, S_2, S_3} p(\theta, s_2, s_3) \left\{ \lg \frac{p(s_3, \theta \mid s_2)}{p(\theta \mid s_2) p(s_3 \mid s_2)} \right\}\\ &\quad = I(\Theta; S_3 \mid S_2). \end{aligned} \tag{16}$$

Please note that for the derivations in (16), we assume that $p(S_3 \mid \Theta, S_1, S_2) = p(S_3 \mid \Theta, S_2)$, i.e., S_3 and S_1 are conditionally independent given both Θ and S_2, where Θ is a random variable representing the fusion result, and S_i is a sensor. The typical relationships between Θ and S_i are illustrated in Fig. 1, where Θ is typically the root node, and S_i's are the leaf nodes in the BN. Given this understanding, if the BN is such that the path (undirected path) between two sensor nodes (e.g., S_1 and S_3) goes through Θ node (e.g., the BN in Fig. 1), then following the D-separation principle for BN, $p(S_3|\Theta, S_1, S_2) = p(S_3|\Theta, S_2)$ holds. Please note that this assumption only holds for some BNs, such as the one in Fig. 1 and the naive BN. It may not hold for an arbitrary BN.

From the chain rule of mutual information, we have

$$I(\Theta; S_3, S_2) = I(\Theta; S_2) + I(\Theta; S_3 \mid S_2). \tag{17}$$

Hence, combining (16) and (17) yields

$$I(\Theta; S_3 \mid S_1, S_2) = I(\Theta; S_2, S_3) - I(\Theta; S_2). \tag{18}$$

Now, we want to apply the similar algebraic process to prove $I(\Theta; S_4 \mid S_1, S_2, S_3) = I(\Theta; S_4 \mid S_3)$ in (19), shown at the bottom of the page, given the Markov conditions that $P(S_4 \mid S_2, S_3) = P(S_4 \mid S_3)$, $P(S_1 \mid S_3, S_2) = P(S_1 \mid S_2)$, $P(S_4 \mid S_1, S_2) = P(S_4 \mid S_2)$, and $P(S_4 \mid S_1, S_3) = P(S_4 \mid S_3)$. By mutual information chain rule, we have $I(\Theta; S_3, S_4) = I(\Theta; S_3) + I(\Theta; S_4 \mid S_3)$, i.e.,

$$I(\Theta; S_4 \mid S_3) = I(\Theta; S_3, S_4) - I(\Theta; S_3). \quad (20)$$

Combining (19) and (20) produces

$$\begin{aligned} I(\Theta; S_4 \mid S_1, S_2, S_3) &= I(\Theta; S_4 \mid S_3) \\ &= I(\Theta; S_3, S_4) - I(\Theta; S_3). \quad (21) \end{aligned}$$

Finally, we can generalize the above process to prove

$$\begin{aligned} &I(\Theta, S_m \mid S_1, S_2, \ldots, S_{m-1}) \\ &\quad = I(\Theta; S_{m-1}, S_m) - I(\Theta; S_{m-1}). \quad (22) \end{aligned}$$

Substituting the results in (18), (21), and (22) into (14) yields

$$\begin{aligned} &I(\Theta; S_1, S_2, \ldots, S_m) \\ &= I(\Theta; S_1) + I(\Theta; S_2 \mid S_1) + I(\Theta; S_3, S_2) \\ &\quad - I(\Theta; S_2) + I(\Theta; S_3, S_4) - I(\Theta; S_3) + \cdots \\ &\quad + I(\Theta; S_{m-1}, S_m) - I(\Theta; S_{m-1}) \\ &= I(\Theta; S_1) + I(\Theta; S_2, S_1) - I(\Theta; S_1) + I(\Theta; S_3, S_2) \\ &\quad - I(\Theta; S_2) + I(\Theta; S_3, S_4) - I(\Theta; S_3) + \cdots \\ &\quad + I(\Theta; S_{m-1}, S_m) - I(\Theta; S_{m-1}) \\ &= I(\Theta; S_1) + \sum_{i=2}^{M} I(\Theta; S_{m-1}, S_m) - I(\Theta; S_{m-1}). \quad (23) \end{aligned}$$

This completes the proof for Theorem 1. ■

B. Proof of Theorem 2

Proof: We want to prove

$$I(\Theta; S_1, \ldots, S_m) \leq I^M(\Theta; S_1, \ldots, S_m). \quad (24)$$

From the mutual information chain rule, we have

$$\begin{aligned} &I(\Theta; S_1, S_2, \ldots, S_m) \\ &\quad = I(\Theta; S_1) + I(\Theta; S_2 \mid S_1) \\ &\qquad + I(\Theta; S_3 \mid S_1, S_2) + I(\Theta; S_4 \mid S_1, S_2, S_3) + \cdots \\ &\qquad + I(\Theta; S_m \mid S_1, S_2, \ldots, S_{m-1}). \quad (25) \end{aligned}$$

By Theorem 1, we have

$$\begin{aligned} &I^M(\Theta; S_1, \ldots, S_m) \\ &\quad = I(\Theta; S_1) + I(\Theta; S_2 \mid S_1) + I(\Theta; S_3 \mid S_2) \\ &\qquad + I(\Theta; S_4 \mid S_3) + \cdots + I(\Theta, S_m \mid S_{m-1}). \quad (26) \end{aligned}$$

By the definition of mutual information, we have

$$I(\Theta; S_3; S_2; S_1) = I(\Theta; S_3; S_2) - I(\Theta; S_3; S_2 \mid S_1) \quad (27)$$

which readily leads to

$$I(\Theta; S_3; S_2 \mid S_1) = I(\Theta; S_3 \mid S_2) - I(\Theta; S_3 \mid S_2, S_1). \quad (28)$$

Hence

$$I(\Theta; S_3 \mid S_2, S_1) = I(\Theta; S_3 \mid S_2) - I(\Theta; S_3; S_2 \mid S_1). \quad (29)$$

Therefore

$$I(\Theta; S_3 \mid S_2) \geq I(\Theta; S_3 \mid S_1, S_2).$$

Please note that we assume here that $I(\Theta; S_3; S_2 \mid S_1) > 0$, which is correct since for our application Θ (the hypothesis)

$$\begin{aligned} &I(\Theta; S_4 \mid S_1, S_2, S_3) \\ &\quad = \sum_{\Theta, S_1, S_2, S_3, S_4} p(\theta, s_1, s_2, s_3, s_4) \left\{ \lg \frac{p(\theta, s_4 | s_1, s_2, s_3)}{p(\theta \mid s_1, s_2, s_3) p(s_4 | s_1, s_2, s_3)} \right\} \\ &\quad = \sum_{\Theta, S_1, S_2, S_3, S_4} p(\theta, s_1, s_2, s_3, s_4) \left\{ \lg \frac{p(\theta, s_4, s_1, s_2, s_3)}{p(\theta, s_1, s_2, s_3) p(s_4 | s_3)} \right\} \\ &\quad = \sum_{\Theta, S_1, S_2, S_3, S_4} p(\theta, s_1, s_2, s_3, s_4) \left\{ \lg \frac{p(s_4 | \theta, s_1, s_2, s_3)}{p(s_4 | s_3)} \right\} \\ &\quad = \sum_{\Theta, S_1, S_2, S_3, S_4} p(\theta, s_1, s_2, s_3, s_4) \left\{ \lg \frac{p(s_4 | \theta, s_3)}{p(s_4 | s_3)} \right\} \\ &\quad = \sum_{\Theta, S_3, S_4} p(\theta, s_3, s_4) \left\{ \lg \frac{p(s_4 | \theta, s_3)}{p(s_4 | s_3)} \right\} \\ &\quad = \sum_{\Theta, S_2, S_3} p(\theta, s_3, s_4) \left\{ \lg \frac{p(s_4, \theta \mid s_3)}{p(\theta \mid s_3) p(s_4 | s_3)} \right\} = I(S_4; \Theta \mid S_3) = I(\Theta; S_4 \mid S_3) \quad (19) \end{aligned}$$

and the other variables (sensors) are not independent of each other.

Similarly, we have $I(\Theta; S_4 \mid S_3) \geq I(\Theta; S_4 \mid S_1, S_2, S_3)$ and $I(\Theta; S_m \mid S_{m-1}) \geq I(S_m \mid S_1, S_2, \ldots, S_{m-1})$.

Hence, (26) $\geq$ (25). The equality sign holds when the Markov property between neighbor sensors is true.

Hence, this completes the proof for Theorem 2. ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable and constructive comments, which help to significantly improve this paper.

REFERENCES

[1] Y. Zhang and Q. Ji, "Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 467–472, Apr. 2006.
[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[3] D. E. Hecherman, E. Horvitz, and B. N. Nathwani, "Toward normative expert systems: Part I. The pathfinder project," *Methods Inf. Med.*, vol. 31, no. 2, pp. 90–105, Jun. 1992.
[4] L. van der Gaag and M. Wessels, "Selective evidence gathering for diagnostic belief networks," *AISB Q.*, vol. 86, pp. 23–34, 1993.
[5] S. Dittmer and F. Jenson, "Myopic value of information in influence diagrams," in *Uncertainty Artif. Intell.*, 1997, pp. 142–149.
[6] V. Bayer-Zubek, "Learning diagnostic policies from examples by systematic search," in *Uncertainty Artif. Intell.*, 2004, pp. 27–34.
[7] V. Krishnamurthy, "Algorithms for optimal scheduling and management of hidden Markov model sensors," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1382–1397, Jun. 2002.
[8] E. Ertin, J. W. Fisher, and L. C. Potter, "Maximum mutual information principle for dynamic sensor query problems," in *Proc. Inf. Process. Sens. Netw.*, San Francisco, CA, 2003, pp. 405–416.
[9] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proc. Inf. Process. Sens. Netw.*, Berkeley, CA, 2004, pp. 36–45.
[10] D. Heckerman, E. Horvitz, and B. Middleton, "An approximate nonmyopic computation for value of information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 3, pp. 292–298, Mar. 1993.
[11] A. Krause and C. Guestrin, "Optimal nonmyopic value of information in graphical models—Efficient algorithms and theoretical limits," in *Int. Joint Conf. Artif. Intell.*, 2005, pp. 1339–1345.
[12] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Uncertainty Artif. Intell.*, 2005, pp. 324–333.
[13] W. Liao and Q. Ji, "Efficient active fusion for decision-making via voi approximation," in *21th Nat. Conf. Artif. Intell.*, 2006, pp. 1180–1185.
[14] Y. Zhang and Q. Ji, "Sensor selection for active information fusion," in *20th Nat. Conf. Artif. Intell.*, 2005, pp. 1229–1234.
[15] W. Liao, W. Zhang, and Q. Ji, "A factor tree inference algorithm for Bayesian networks and its application," in *Proc. 16th ICTAI*, Boca Raton, FL, 2004, pp. 652–656.
[16] W. Zhang and Q. Ji, "A factorization approach to evaluating simultaneous influence diagrams," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 4, pp. 746–757, Jul. 2006.
[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.
[18] F. V. Jensen and F. Jensen, "Optimal junction trees," in *Proc. 10th Conf. Uncertainty AI*, San Francisco, CA, 1994, pp. 360–366.
[19] R. Dechter, "Bucket elimination: A unifying framework for probabilistic inference," in *Proc. 12th Conf. Uncertainty AI*, San Francisco, CA, 1996, pp. 211–219.

Yongmian Zhang (M'04) received the Ph.D. degree in computer engineering from the University of Nevada, Reno, in 2004.

He holds a research position with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY. His areas of research include information fusion, computer vision, and human–computer interactions.

Qiang Ji (SM'04) received the Ph.D. degree in electrical engineering from the University of Washington, Seattle.

He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering and the Director of the Intelligent Systems Laboratory, Rensselaer Polytechnic Institute, Troy, NY. He is also a Program Director with the National Science Foundation (NSF), Arlington, VA, where he is responsible for managing part of the NSF's computer vision and machine learning programs. He previously held teaching and research positions with Beckman Institute, University of Illinois, Urbana; Robotics Institute, Carnegie Mellon University, Pittsburgh, PA; the Department of Computer Science, University of Nevada, Reno; and the U.S. Air Force Research Laboratory. His research has been supported by major governmental agencies, including the NSF, the National Institutes of Health, the Defense Advanced Research Projects Agency, the Office of Naval Research, the Army Research Office, and the Air Force Office of Scientific Research, as well as by major companies, including Honda and Boeing. He has published more than 150 papers in peer-reviewed journals and conference proceedings. He is an Editor for several computer-vision- and pattern-recognition-related journals.

Prof. Ji has served as program committee member, area chair, and program chair in numerous international conferences/workshops.