Hindawi Publishing Corporation
Advances in Multimedia
Volume 2008, Article ID 543674, 10 pages
doi:10.1155/2008/543674

Research Article

A Theoretical Framework for Quality-Aware Cross-Layer Optimized Wireless Multimedia Communications

Song Ci,[1] Haohong Wang,[2] and Dalei Wu[1]

[1] *Department of Computer and Electronics Engineering, University of Nebraska-Lincoln, NE 68182, USA*
[2] *Marvell Semiconductors, Santa Clara, CA 95054, USA*

Correspondence should be addressed to Song Ci, sci@engr.unl.edu

Received 20 May 2007; Revised 5 September 2007; Accepted 5 November 2007

Recommended by Jianwei Huang

Although cross-layer has been thought as one of the most effective and efficient ways for multimedia communications over wireless networks and a plethora of research has been done in this area, there is still lacking of a rigorous mathematical model to gain in-depth understanding of cross-layer design tradeoffs, spanning from application layer to physical layer. As a result, many existing cross-layer designs enhance the performance of certain layers at the price of either introducing side effects to the overall system performance or violating the syntax and semantics of the layered network architecture. Therefore, lacking of a rigorous theoretical study makes existing cross-layer designs rely on heuristic approaches which are unable to guarantee sound results efficiently and consistently. In this paper, we attempt to fill this gap and develop a new methodological foundation for cross-layer design in wireless multimedia communications. We first introduce a delay-distortion-driven cross-layer optimization framework which can be solved as a large-scale dynamic programming problem. Then, we present new approximate dynamic programming based on significance measure and sensitivity analysis for high-dimensional nonlinear cross-layer optimization in support of real-time multimedia applications. The major contribution of this paper is to present the first rigorous theoretical modeling for integrated cross-layer control and optimization in wireless multimedia communications, providing design insights into multimedia communications over current wireless networks and throwing light on design optimization of the next-generation wireless multimedia systems and networks.

1. INTRODUCTION

In recent years, ubiquitous computing devices such as laptop computers, PDAs, smart phones, automotive computing devices, and wearable computers have been ever growing in popularity and capability, and people have begun more heavily to rely on these ubiquitous computing devices. Therefore, there has been a strong user demand for bringing multimedia streaming to the devices such as iTunes, PPLive, MSN, and YouTube. However, bringing delay-sensitive and loss-tolerant multimedia services based on the current wireless Internet is a very challenging task due to the fact that the original design goal of the Internet is to offer simple delay-insensitive loss-sensitive data services with little QoS consideration. Therefore, this shift of design goal urges us to rethink the current Internet architecture and develop a new design methodology for multimedia communications over the current and future wireless Internet. So far, cross-layer design has been thought as one of the most effective and efficient ways to provide quality of service (QoS) over wireless networks, and it has been receiving many research efforts. The basic idea of cross-layer design is to fully utilize the interactions among design variables (system parameters) residing in different network functional entities (network layers) to achieve the optimal design performance of time-varying wireless networks.

In order to achieve the global optimality of cross-layer design, we need to consider design variables and the interactions among them as much as possible. However, more does not necessarily mean better. The more design variables we consider, the more difficult is orchestrating a large number of design variables to make them work harmonically and synergetically. From the point of view of nonlinear optimization, the number of design variables increases and the size of state space of the objective function will increase exponentially, making the optimization

problem unmanageable. To overcome this problem, one often used approach is to reduce the size of the problem at the system modeling phase and then solve the simplified problem by using various optimization algorithms such as gradient-based local search, linear/nonlinear programming, genetic algorithm, exhaustive search, and heuristic-based approach like artificial neural networks.

However, reducing a high-dimensional cross-layer optimization problem to a low-dimensional problem in the system modeling phase raises a series of questions:

(1) how to evaluate the fidelity of the simplified problem compared with the problem as what it should be,

(2) how to evaluate the quality of the suboptimal solution to the global optimum,

(3) how to evaluate the robustness of the solution, that is, whether the solution can guarantee the predictable sound results at all possible circumstances.

Unfortunately, at the time of this writing, we have no clear answers to all these three questions.

Moreover, reducing the size of the problem in the problem formulation means that only part of the current Internet architecture can be considered, causing a shift of the design goal of multimedia services from the best user experience to some layer-specific performance metrics such as distortion at the application layer, delay at the network layer, and goodput at the MAC/PHY layer. This shift of design goal may cause an "Ellsberg paradox," where each individual design variable makes good decisions for maximizing the objective function. But the overall outcome violates the expected utility function. In other words, breaking a big problem into several smaller problems in the system modeling phase can only increase the solvability of the original problem but cannot guarantee that it is a good solution. The "Ellsberg paradox" also tells us that the traditional additive measure such as probability measure may no longer hold in the context of cross-layer design due to the possible strong coupling (interdependency) among design variables. At the point of this writing, there have been many researches done on interdependency modeling in the context of cross-layer design, but they are mostly qualitative rather than quantitative approaches, and their applications are still within the scope of local cross-layer optimization.

We argue that all aforementioned difficulties in the area of cross-layer design of wireless multimedia communications are due to lacking of methodological foundation and in-depth understanding of cross-layer behavior. Our goal is to provide a flexible yet scalable theoretical cross-layer framework to accommodate all major design variables of interest, spanning from application layer to physical layer, for delay-bounded multimedia communications over wireless single/multihop networks. We start from proposing an integrated cross-layer framework for the best user experience. Although the engineering side of cross-layer design is not the main focus of this paper, we still briefly discuss how to utilize the methodological foundation to achieve real-time multimedia communications through a fast algorithm for large-scale global cross-layer optimization based on quantitative significance measure and sensitivity analysis.

The rest of the paper is organized as follows. We briefly introduce the related work in Section 2. In Section 3, we present a unified theoretical cross-layer framework for wireless multimedia communications based on link adaptation, rate-distortion theory, and dynamic programming. A further discussion of how to apply the proposed methodological foundation for real-time applications is made in Section 4, where new feature-based approximate dynamic programming is introduced, followed by the conclusion in Section 5.

2. RELATED WORK

In literature, topics involving video delivery over multihop networks such as video coding, multihop routing, QoS provisioning, link adaptation are separately studied. Therefore, the corresponding video compression efficiency and the transmission efficiency are also separately optimized. In prediction mode, selection of video coding, periodic intracoding of whole frames [1], continuous blocks [2], or random blocks [3] has been firstly proposed. These methods apply intracoding uniformly to all the regions of the frame. Then, "content-adaptive" methods are proposed to apply frequent intra-update to regions that undergo significant changes [4], or where a rough estimate of decoder error exceeds a given threshold [5, 6]. A significant advance over the above early heuristic mode switching strategies is the rate-distortion (RD) optimized mode selection. The RD optimized mode selection is achieved by choosing a mode that minimizes the quantization distortion between the original frame/macroblock and the reconstructed one under a given bit budget [7, 8]. However, the encoders in [7, 8] have no capability to accurately estimate the overall distortion. So, the selected prediction mode is not necessarily optimal. The work in [9] proposes an algorithm to optimally estimate the overall distortion of decoder frame reconstruction due to quantization, error propagation, and error concealment. The accurate estimate is integrated into a rate-distortion-based framework for optimal switching between intracoding and intercoding modes per macroblock. However, the joint optimization between mode selection and video transmission parameters under wireless environment is not addressed in [9]. The work in [10] presents an end-to-end approach to solve the fundamental problem of RD optimized mode selection over packet-switched networks, but it only aims at Internet peer-to-peer video communication.

In routing for video delivery in multihop networks, an application-centric cross-layer approach has been proposed to formulate an optimal routing problem for multiple description video communications [11]. Physical and MAC layer dynamics of wireless links are translated into network layer parameters. The application layer performance, that is, average video distortion, is considered as the function of network layer performance metrics, for example, bandwidth, loss, and path correlation. But the routing metric, that is, average video distortion, is roughly computed from a simple rate-distortion model without discussion on selection of source coding parameters. The same problem goes to [12]

and [13] even though optimal paths are selected optimizing the quality under various constraints. In addition, in [13] exhaustive algorithm is adopted for the determination of the cross-layer optimized mesh-network path selection, which may incur heavy computational load and make it unpractical for real applications.

Cross-layer optimized wireless video has been studied from different aspects, such as cross-layer architecture [14, 15], content analysis [16–18], video compression and RD optimization [2–4, 6–10, 19–21], source packetization [22, 23], QoS provisioning [24–26], application-centric routing [11–13, 27], queueing and scheduling [28–31], energy efficiency [32, 33], and link adaptation [34, 35]. To reach a global optimality at the level of frame or video sequence rather than at the level of packet, we need to evaluate the overall distortion and the effect of packet pipelining in a network on the total delay of a frame or a video sequence. To the best of our knowledge, although some works focus on the cross-layer design for video delivery over multihop wireless networks, there is still no substantial work that can reach such kinds of global optimality.

In wireless video, optimization has to be done over multiple source coding units, such as frames and pixel blocks, for the best reconstructed video quality. There is "*only one exact method for solving problems of optimization over time; in the general case of nonlinearities with random disturbance, it is dynamic programming (DP)*" [36]. However, the biggest challenge of applying dynamic programming in practical large-scale problems is *curse of dimensionality* [37], where the size of state space normally increases exponentially with the number of control variables increasing. Therefore, the most sensible way is to map a huge state space $\Re^n$ to a much smaller feature space $\Re^m$ ($m \ll n$), which is called *approximate dynamic programming (ADP)*, also known as neurodynamic programming, adaptive dynamic programming, adaptive critics, or reinforced learning, depending on in which discipline the technique is used [36, 38, 39].

Existing ADP approaches have largely ignored the interdependencies among control variables, which might lead to *loose approximation error bounds*. Nonadditive measure theory was developed to characterize the interactions among control variables [40–43], and it has been widely used in various areas. Choquet integral [44] is regarded as the most effective and efficient way to calculate nonadditive measure and has received a significant amount of research [45–48]. Since nonadditive measure is defined on the power set, fast algorithms [49] have been studied to speed up the calculation process. However, current research on nonadditive measure still focuses on static linear systems with commensurable data [50].

3. A THEORETICAL CROSS-LAYER FRAMEWORK FOR WIRELESS MULTIMEDIA COMMUNICATIONS

3.1. Problem statement

In the protocol stack of multimedia over wireless networks, each layer has one or multiple key system parameters which would significantly impact the overall system performance. At the application layer, tradeoff between rate and distortion is an inherent feature of every lossy compression scheme for video source coding. Prediction mode and quantization level are two critical parameters. At the network layer, routing algorithm is important to find the best delivery path over a single/multihop wireless network. At the data link layer, hybrid automatic repeat request (HARQ), media access control protocols, and packetization are often used to maintain a low packet loss rate. However, the choice of maximum retransmission number is a tradeoff between resultant packet delay and packet loss rate. Note that for real-time multimedia applications, we might not consider HARQ due to strict delay constraints. At the physical layer, adaptive modulation and coding scheme is an important tradeoff between transmission rate and packet loss rate. Furthermore, the end-to-end performance is not completely determined by the parameters of individual layer, but rather by all parameters of all layers. For example, the end-to-end delay consists of propagation delay (determined by the number of hops of the selected path), transmission delay (determined by channel conditions, modulation and channel coding, maximum retransmission number, and source rate), and queueing delay (determined by source rate, transmission rate, and the selected path). Moreover, due to the time-varying nature of wireless channels, each node in the network should be capable of adjusting these parameters quickly to maintain a good instantaneous performance. Clearly, the layer-separated design no longer guarantees an optimal end-to-end performance for multimedia delivery over wireless networks.

3.2. Methodology

We develop a cross-layer framework to optimize multimedia communications over single/multihop wireless networks. In order to demonstrate the main idea of the proposed framework as shown in Figure 1, at the application layer, we implement our framework based on the ITU-T H.264 standard. The rate-distortion tradeoff in video source coding makes it very critical to select suitable video coding parameters such as prediction mode (PM) and quantization parameter (QP). Without losing generality, we consider a multihop wireless network scenario in which all nodes can act as either a source or destination as well as a router for other nodes. To carry out end-to-end delay-bounded multimedia communications, at the network layer, we assume that certain routing protocols are used to come up with the routing table. Then, a quality-aware routing algorithm needs to be developed to select the best multihop path from the source to the destination. Each hop adopts adaptive modulation and coding (AMC) at the physical layer to overcome the adverse effects caused by the time-varying channel condition.

Let us denote by W the number of frames of a video clip, $f_1, f_2, \ldots, f_W$, and let $m_i^1, m_i^2, \ldots, m_i^M$ be the macroblocks of frame f_i. Since each frame is processed in units of macroblock (corresponding to $16*16$ pixels in the original frame), let v_i^j denote the coding parameter vector

of macroblock j in frame i as quantization parameter (QP) and prediction mode choice (I or P frame). Let $B_i^j(v_i^j)$ denote the consumed bits in coding the macroblock m_i^j with the coding parameter vector v_i^j; then the total bits consumed by the frame can be expressed as $F_i = \sum_{j=1}^{M} B_i^j(v_i^j)$.

We assume that the considered multihop network consists of Z nodes $\{N_1, N_2, \ldots, N_z\}$. For any two nodes N_x and N_y, if N_x can directly communicate with N_y, we say that there exists a hop between N_x and N_y. Let $l_i(x \rightarrow y)$ denote the hop between the node N_x and the node N_y. Considering the time-varying nature of the network, let $L^i(x \rightarrow y) | 1 \le x \le Z, 1 \le y \le Z, x \ne y$, denote all the connectivity information within the network when transmitting the frame f_i. Accurate L^i can be obtained from certain routing protocols such as OLSR routing protocol. Let $P^i = \{l_1, l_2, \ldots, l_G\}$ be a path $\{l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_G\}$ for transmitting frame f_i from the source node to the destination node. Clearly, there exist $P^i \subseteq L^i$ and $1 \le G_i \le Z - 1$. Let us denote $\{\gamma_i^p\}$ and $\{R_i^p\}$ as the channel SNR and transmission rate of the link l_p with $1 \le p \le G$, $\{A_i^p\}$ and C_i^p as the modulation mode and associated channel coding rate, and $N_{R_i}^p$ as the number of retransmissions. Then, the delay in transmitting the frame f_i on the link l_p can be written as $\{T_i^p(A_i^p, C_i^p, N_{R_i}^p, \gamma_i^p, R_i^p, F_i)\}$. Clearly, the total delay in transmitting the whole video clip can be expressed by

$$T = \sum_{i=1}^{W} \sum_{p=1}^{G_i} T_i^p(A_i^p, C_i^p, N_i^p, \gamma_i^p, R_i^p, F_i). \tag{1}$$

Let $\widetilde{f}_i$ denote the reconstructed ith frame at the receiver side. Using the mean square error as distortion metric, the overall expected distortion for the whole video clip is

$$E[D] = \sum_{i=1}^{W} E[d(f_i, \widetilde{f}_i)]. \tag{2}$$

Note that in this work, $d(\cdot)$ of $E[D]$ can be calculated by any distortion estimation method such as the mean square error (MSE) estimation method and the recursive optimal per-pixel estimate (ROPE) method. Likewise, any error concealment schemes can be used at the receiver side to further enhance the perceivable video quality. Since the formulation discussed above considers W consecutive video frames, the spatial-temporal correlation among frames and macroblocks has been taken into account in the global optimization framework.

Thus, the proposed cross-layer framework for wireless multimedia communications can be formulated as

$$\text{Min}\, E[D], s.t. : T \le T_{\max}, \tag{3}$$

where $T_{\max}$ is a predefined delay budget for delivering the given video clip.

Recall that the focus of the proposed framework is to jointly find the optimal parameter set for each frame f_i, including the source coding v_i^j, the delivery path P^i, the maximum number of retransmissions N_i^p, and the modulation A_i^p with the associate coding C_i^p. Here, p is the index of each hop on the path P^i. Clearly, the optimal solution for the problem described by (3) can be written as

$$\min_{\{v,P,N_R,A,C\}} \sum_{i=1}^{W} E[d(f_i, \widetilde{f}_i)] \tag{4}$$

with the delay constraint

$$\sum_{i=1}^{W} \sum_{p=1}^{G_i} T_i^p(A_i^p, C_i^p, N_i^p, \gamma_i^p, R_i^p, F_i) \le T_{\max}. \tag{5}$$

Clearly, in (4) we assume that the decoder side has a sufficient size buffer to hold part of the decoded video frames, say, a group of pictures. Given the dramatically fast growing silicon performance and the decreasing size and cost for the memory and silicon, the assumption is reasonable for most scenarios. But when the size of decoder buffer is constrained, (4) would be rewritten as follows:

$$\min_{\{v,P,N_R,A,C\}} E[d(f, \widetilde{f})] \tag{6}$$

with the delay constraint

$$\sum_{p=1}^{G} T^p(A^p, C^p, N^p, \gamma^p, R^p, F) \le T_{\max}, \tag{7}$$

where f represents each of the W frames, which has a delay constraint. Clearly, (6) does add difficulties on top of (4), although a number of constraints are included to eliminate some valid solutions for the original problem.

Note that the unique feature of (4) is that it is essentially a convex function, which has been shown in large amount of research done on rate-distortion relationship under the context of multimedia processing and transmissions. In other words, there always exists a global optimality of this formulation. This is a very important conclusion, since other existing global cross-layer optimization frameworks focusing on network QoS or using decomposition approach cannot guarantee the convexity of all decomposed subproblems. For a given multimedia application, the global optimization problem described in (4) turns into a constrained nonlinear optimization problem, which can be solved by *Lagrangian multipliers (LMs)* and *Lagrangian relaxation (LR)* [51]. So, we can use the derived Lagrangian cost function as the unified cost function. In this work, the cost function J is the average distortion over the given video clip $E(D)$.

For the global optimality of system performance, we need to optimize current control action u_t over time $t + 1, t + 2, \ldots, N$; in other words, current control action $u_t = \{v, P, N_R, A, C\}$ needs to be chosen with considerations of future cost J. For example, the end user will evaluate the perceivable video quality based on the overall quality of the whole video clip rather than the quality of each individual

video frame. Therefore, the cost function for optimization over time based on (4) is

$$\arg\min_{u_t \in U} C_i^{u_t}(x_t) + J(x_{t+1}) \mid x_t. \tag{8}$$

Here, x_t is the state at time t, and the value $J(x_{t+1})$ is introduced to capture the future cost (i.e., $E(D)$) at time $t+1$ incurred as a result of taking the control action u_t at time t.

So far, there is only one exact method for global optimization over time with nonlinearities and random disturbances [36], which is *dynamic programming (DP)*. DP provides methods for choosing a value function $J(\cdot)$ to derive an optimal policy $\pi = \{u_0, u_1, u_2, \ldots, u_{N-1}\}$. There has been a plenty of research on how to use DP-based algorithms for multimedia processing and transmission. In order to use DP to find the global optimality of (4), a unified cost-to-go function J has to be constructed:

$$J^{u_t}(x_t) = \min_{u_t,\ldots,u_{N-1}} \sum_{i=t}^{N-1} J(x_i, u_i(x_i), x_{i+1}) | x_t. \tag{9}$$

Then, the global optimization problem turns into calculating the cost-to-go function $J_0(x_0)$, which is the overall cost to be incurred in the finite horizon of N steps.

3.3. Numerical results

We have evaluated the performance of the proposed integrated cross-layer framework through extensive simulations based on H.264 JM12.2 codec. In general, we are interested in comparing our integrated cross-layer design with the best possible results of H.264 codec. Our goal here is to illustrate the difference of performance gain between the global optimality achieved by the proposed framework and the superposition of multiple local optimality done separately at different network layer (s). In this paper, the best baseline performance is derived: (1) at the application layer, it uses the rate control scheme of H.264 codec; (2) at the network layer, it always chooses the path with the best average SNR at each hop; (3) at the MAC and PHY layers, it always chooses the AMC scheme for the shortest delay while keeping the predefined PER performance.

From the simulation results, up to 3 dB PSNR gain can be achieved by using the proposed approach compared with using the existing piecemeal approach, as shown in Figure 2.

Remark 1. We have proposed a top-down theoretical cross-layer framework for multimedia over wireless networks, and the correctness of the proposed methodology is based on its rigorous theoretical foundation. Moreover, the proposed methodology is based on dynamic programming, which means that it is very flexible and scalable; any interaction of interest in the system can be easily integrated into the proposed framework. Since we consider all the major interactions of interest spanning from application layer to physical layer, we have overcome the major drawback of existing cross-layer designs where the simplification occurs at the system modeling phase rather than the problem solving phase. Therefore, the proposed methodology provides the true global optimality and a new design guidance to the cross-layer design for multimedia over various wireless networks.

4. FURTHER DISCUSSION

In this section, we will further discuss how to apply the aforementioned global optimization framework for real-time multimedia communications as formulated in (4). This is not only practically important but also theoretically interesting.

4.1. Problem statement

So far, we have presented a new theoretical framework for cross-layer design of multimedia communications over wireless networks, which provides a sound methodological foundation for us to evaluate cross-layer designs using dynamic programming (DP) which has been widely adopted to study sequential decision-making problems (stochastic control). However, the practical applications of dynamic programming are limited mostly due to the dual curses of dimensionality and uncertainty, that is, the large size of underlying state space of the *cost-to-go function* which is a function of the current state for evaluating the expected future cost to be incurred. The "curse of dimensionality" means that the computational complexity of the cross-layer design can be increased exponentially when the number of considered design variables increases. The "curse of uncertainty" (modeling) indicates the fact that in a complex networking system there exist various uncertainties making it very difficult to know the explicit system model and/or states. Generally speaking, uncertainties can be classified into two categories: measurement uncertainties and model uncertainties. Under the context of cross-layer design, measurement uncertainties are mainly caused by randomness in data collecting process such as inaccurate channel feedback, while model uncertainties are mainly caused by various approximations made in system modeling process such as approximations made on channel quality, traffic load, node mobility, number of users, and user behaviors. For cross-layer design, uncertainties existing in interdependency among design variables may cause severe performance degradation. Therefore, the "dual curses" make cross-layer optimization a very challenging problem.

4.2. Methodology

4.2.1. Feature-based approximate dynamic programming

The most sensible and rational way to deal with the difficulty caused by "dual curses" is to generate a compact parametric representation (compact representation, for brevity) to approximate the cost-to-go function for a significant complexity reduction through mapping the huge state space to a much smaller feature space characterized by a compact representation.

Currently, the selection of a compact representation largely relies on heuristics which somewhat contradicts

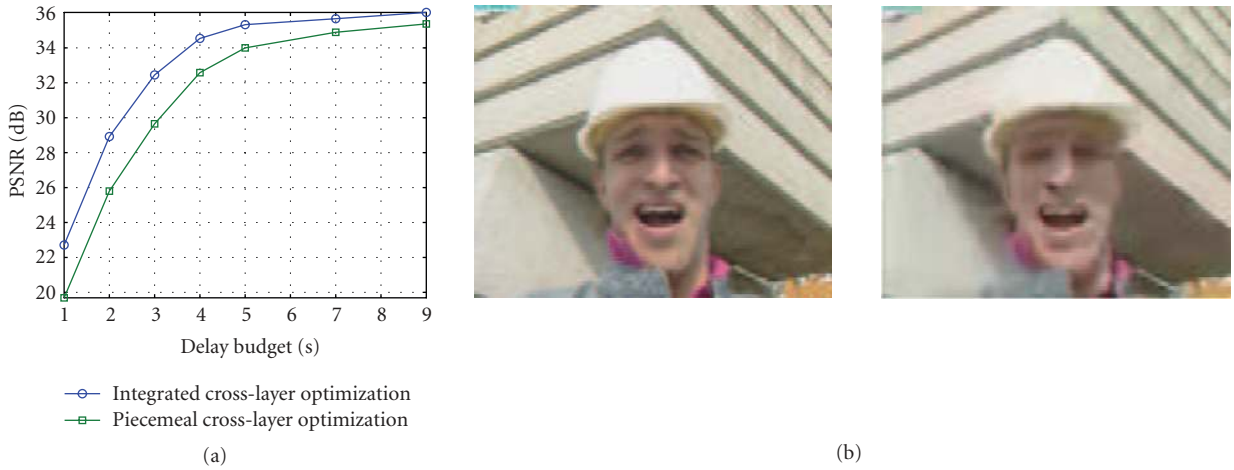

Figure 1: (a) Performance comparison using sample video clip: global cross-layer optimization versus existing piecemeal cross-layer optimization. Here, assume that multihop paths and their link quality can be found by a multihop routing protocol, such as optimized link state routing protocol (OLSR) [52]. In this simulation, the average link SNRs (in dB) of three multihop paths are $P_1 = \{5, 10, 15\}$, $P_2 = \{5, 10, 20\}$, and $P_3 = \{5, 15, 25\}$. Six AMC schemes as listed in [51] are adopted at the PHY layer. At the receiver side, a simple error concealment algorithm is adopted where the lost macroblock will be replaced by the latest correctly received one. (b) Perceptual video quality comparison based on H.264 codec with the same delay budget, where (a) is global optimality and (b) is the best baseline.

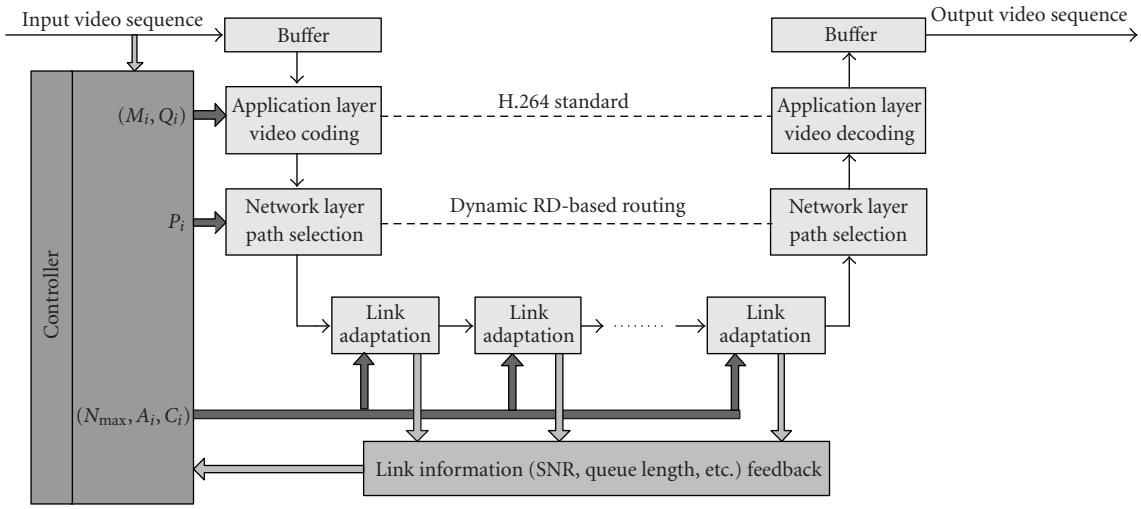

Figure 2: An integrated cross-layer framework of multimedia communications over multihop wireless networks.

Figure 3: An integrated cross-layer framework of multimedia communications over multihop wireless networks.

Table 1: Perceptual video quality comparison based on H.264 codec in a real-time environment $T_{\text{frame}} = 35\,ms$), where (a) globally optimal, (b) near optimal, and (c) the best baseline.

u[1]	QP	179.23	T	Global	$ADP - P_1$	$ADP - P_2$	$ADP - P_3$	Baseline
u[2]	Path	31.95	0.01312	146433	164824	156336	149483	194813
u[3]	QP,Path	93.31	0.02217	88552	98207	97470	95454	123607
u[4]	AMC	23.18	0.03454	56983	67341	64555	60364	78341
u[5] QP,AMC 252.49	0.03960	49719	54891	51555	50201	63891		
u[6]	Path,AMC	3.22	0.06105	33702	36535	35623	34564	42535
u[7]	QP,Path,AMC	157.31	0.09725	21573	23551	23337	22811	27551

the nonheuristic aspects of the dynamic programming methodology. Therefore, we propose a new method based on nonadditive measure theory, which can dynamically generate compact representations of the huge state space. Unlike other nonlinear feature-extraction approaches such as artificial neural network, the proposed method is adaptive and nonheuristic in the sense that it allows us to quantitatively characterize the significance or the desirability of state vectors with considerations of interactions among different state variables. Therefore, new feature-based approximate dynamic programming can be developed based on the adaptive feature extraction and compact representation.

We consider a large-scale dynamic programming problem defined on a finite state space S. Let n denote the cardinality of S; thus we have $S = 1, \ldots, n$, and $n = \Pi_{k=1}^{\omega} N_k$, where N_k is the number of control actions for the parameter $k \in [1, \omega]$. Our goal is to quantitatively characterize the significance effect of parameters on the cost-to-go function J.

4.2.2. Feature-based compact representation

In the context of dynamic programming, the cost-to-go vector J is defined as a vector whose components are the cost-to-go values of various states. The cost-to-go function specifies the mapping from states to cost-to-go values. Therefore, the optimal cost-to-go vector J^* of policy π with initial state i is defined by

$$J_i^{\pi^*} = \min_{\pi} J_i^{\pi}, \quad i \in S, \tag{10}$$

and the policy at state i is defined by

$$\pi_i^* = \arg\min_{u \in U(i)} \left(C_{iu} + \sum_{j \in S} J_j^* \right), \quad \forall i \in S. \tag{11}$$

The dynamic programming problem is to seek the optimal policy π^* to achieve

$$J_i^{\pi^*} = J_i^*, \quad \forall i \in S. \tag{12}$$

In large-scale dynamic programming problems, the size of state space normally increases exponentially with the number of state variables, making it extremely difficult to compute and store each component of the cost-to-go function. Therefore, the most sensible way is to map a huge state space $\Re^n$ to a much smaller feature space $\Re^m$ ($m \ll n$). Formally, a compact representation can be described as a scheme for recording a high-dimensional cost-to-go vector $V \in \Re^n$ using a lower-dimensional parameter vector $W \in \Re^m$. So, if we can obtain an approximation of $J \in \Re^m$ to $J^* \in \Re^n$, we may still generate a near optimal control policy π_J but with significant computational acceleration satisfying

$$\pi_J = \arg\min_{u \in U(i)} \left(C_{iu} + \sum_{j \in S} J_j \right), \quad \forall i \in S. \tag{13}$$

In the context of approximate dynamic programming, we would like to see that when J approaches J^*, π_J is getting close to π^*. Therefore, a compact representation can be described as a mapping of $J : \Re^n \mapsto \Re^m$ to $W :\in \Re^m$ associated with a cost-to-go vector. Each component of $\widetilde{J}_i(W)$ of the mapping is the i th component of a cost-to-go vector represented by the parameter vector W.

Formally, a feature f is defined as a function from the state space S into a finite set Q of feature values. In stochastic multistage decision processes, we might need several features, $f_1, f_2, \ldots, f_K$, forming a feature vector $F(i) = (f_1(i), \ldots, f_K(i))$ for each state $i \in S$. The feature vector $F(i)$ indicates the desirability or significance of the associated state i. Therefore, for a feature-based compact representation, the component $\widetilde{J}_i(W)$ of $\widetilde{J}(W)$ can be written as $\widetilde{J}_i(F(i), W)$.

For approximate dynamic programming using feature-based compact representation, the approximate cost-to-go function is

$$\widetilde{J}(W) = g(F(i), W), \tag{14}$$

where g is defined as an approximation architecture $g : \Re^K \times \Re^m \mapsto \Re$ with $\Re \in \Re^n$, meaning that g will only cover the most significant finite region of $\Re^n$. In order to achieve the best quality of approximation, it would be highly desirable to have effective and efficient parameter-selection and feature-extraction algorithms. Unfortunately, the existing feature-extraction and parameter-selection algorithms are mainly based on heuristics such as Q-learning and neural network, but those methods lack for sound engineering judgement.

4.2.3. Feature extraction and parameter selection based on significance measure

Feature extraction requires us to catch the "dominant nonlinearities" in the optimal cost-to-go function J^*. Then,

based on the extracted features, the parameter vectors W can be determined and so can be $\Re^m$.

In our preliminary study [53], a new method for feature extraction, called *significance measure*, has been proposed based on nonadditive measure theory [40]. The unique feature of significance measure is that the nonlinear interactions among state variables on the cost-to-go function can be quantitatively measured by solving a generalized nonlinear Choquet integral. As shown in our preliminary study [53], the feature-based approximation can be expressed as

$$\Delta\widetilde{J} = \sum_{k=1}^{m} \Delta f(x) \cdot \mu_k + \xi, \tag{15}$$

where x is state variable, J is the cost-to-go function, and $f(x)$ is observation of state variable. The impact of interactions among state variables on the cost-to-go function is described by a *set function* μ defined on the power set of state variables satisfying the condition of vanishing at the empty set, that is, $\mu : P(X) \to (-\infty, +\infty)$ with $\mu(\varnothing) = 0$. The set function μ is called nonadditive measure [40]. There has been a lot of research done to find the optimal μ by solving the nonlinear integral equation such as Choquet integral [48, 54] based on a set of observation data. An advantage of the proposed significance measure method described above is that it only needs system operation data (simulations), which can be easily acquired from the device drivers. Therefore, it is fairly efficient in terms of computation and storage. Significant measure and sensitivity analysis.

Once, we determine the significance measure of state variables $\mu_1, \mu_2, \ldots, \mu_{2^\omega - 1}$ corresponding to different parameter sets. Then, the parameter set with the largest μ_i can be directly used for parameter selection. Furthermore, the value of each parameter set can be interpreted as feature, since it reflects the parameter significance towards the cost-to-go function. We can choose the parameter set having the largest value of μ_i to be the compact representation $W \in \Re^m$ of the high-dimensional cost-to-go vector $V \in \Re^n$. Therefore, various approximate dynamic programming approaches using feature-based compact representation can utilize the new method for compact representation, feature extraction, and parameter selection. For example, if we adopt feature-based look-up table approximate dynamic programming architecture, the approximated cost-to-go function is $\widetilde{J}_i(W) = W$, or we can use $\widetilde{J}_i(W) = W^T F(i)$ if using linear approximate architecture.

4.3. Numerical results

As discussed earlier, based on the significance measure and sensitivity analysis, we can derive a new method for feature extraction and compact representation for approximating the original large-scale dynamic programming. Using the same problem setting as of Figure 1, a simple example to illustrate the basic idea of the proposed approach is devised. First, an operational data set in the format of *[QP, Path, AMC, Value of cost-to-go function]* has been collected by uniformly sampling the dynamic programming state space. Then, the significance measure algorithm, as presented in [53, 55] was applied to the collected data. The derived significance measure of control variables and their interdependencies can be derived as shown in Tabl @ IV-B3, where columns 1–3 represent significance measure of control variables, where u (column 1) indicates the significant impact of each subset of control variables (QP, $Path$, AMC) (column 2) on the cost-to-go function (column 3) based on the collected measurements. The original three-dimension (QP, AMC, $Path$) DP problem can be approximated by a two-dimension (QP, AMC) ADP problem. Columns 4-9 represent MSE distortion of DP versus ADP versus the best baseline under different frame delay budgets (T), where three ADP values are corresponding to adopt different fixed paths (P_1, P_2, or P_3) in the approximation.

In this simulation, based on the significance measure, the interaction between QP and AMC has the most significant impact on the cost-to-go function, meaning that "path" is not as significant as the other variables. So, it could be excluded from the optimal search. This way, the cardinality of the approximated state space can be *reduced by three times*. Compared with the global optimal performance, the maximum approximation error caused by excluding path from the DP search is 12.5%, corresponding to the shortest delay budget; however, in this case, the result of ADP-based solution still outperforms the best baseline H.264 performance by 15.4%.

Remark 2. In this section, we propose a new method for feature extraction and compact representation of approximate dynamic programming, which is based on the significance measure of each set of design variables. We discuss a novel feature-based approximate dynamic programming approach for solving the large-scale dynamic programming problem in support of real-time multimedia applications. Furthermore, since all the significant measures of a power set of design variables are available, a scalable complexity framework by exploring the tradeoff between the quality of approximation (QoA) and the quality of service (QoS) could be developed in future. Note that the proposed significance measure method and the feature-based approximate dynamic programming approach are fairly generic and are applicable for any large-scale design optimization and real-time control scenarios.

5. CONCLUSION

The major challenges of current cross-layer design for multimedia communications over wireless networks are (1) lacking of understanding of cross-layer behaviors, (2) simplifying cross-layer design at the system modeling phase, and (3) relying on heuristic approaches. We argue that all these challenges are caused by lacking of a new methodology for cross-layer design of multimedia communications over wireless networks. This has motivated us to propose a new methodological foundation for cross-layer design of multimedia communications over wireless networks, which has made two major contributions to the research area: (1) the theoretical framework with major design variables spanning from application layer to physical layer for cross-layer design of multimedia communications over wireless

networks, and (2) the novel feature-based approximate dynamic programming approach based on a new significance measure method to understand cross-layer behaviors and speed up large-scale cross-layer optimization. The proposed methodological foundation is fairly general and can be applicable to other applications in multimedia communications. However, we are *not* trying to solve all the problems in this paper; rather, we are trying to look into this challenging problem from a different angle and open up a new research direction for future studies in the field of wireless multimedia communications. We believe that the proposed methodological foundation will significantly contribute to the emerging research areas such as service- and application-oriented QoS provisioning in the future Internet.

REFERENCES

[1] T. Turletti and C. Huitema, "Videoconferencing on the internet," *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 340–351, 1996.

[2] Q.-F. Zhu and L. Kerofsky, "Joint source coding, transport processing, and error concealment for H.323-based packet video," in *Visual Communications and Image Processing (VCIP '99)*, vol. 3653 of *Proceedings of SPIE*, pp. 52–62, San Jose, Calif, USA, January 1999.

[3] G. Côté and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the Internet," *Signal Processing: Image Communication*, vol. 15, no. 1-2, pp. 25–34, 1999.

[4] P. Haskell and D. Messerschmitt, "Resynchronization of motion compensated video affected by ATM cell loss," in *Proceedings of IEEE International Conference on Speech, Acoustics and Signal Processing (ICASSP '92)*, vol. 3, pp. 545–548, San Francisco, Calif, USA, March 1992.

[5] J. Y. Liao and J. D. Villasenor, "Adaptive intra update for video coding over noisy channels," in *Proceedings of IEEE International Conference on Image Processing (ICIP '96)*, vol. 3, pp. 763–766, Lausanne, Switzerland, September 1996.

[6] E. Steinbach, N. Färber, and B. Girod, "Standard compatible extension of H.263 for robust video transmission in mobile environments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 6, pp. 872–881, 1997.

[7] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[8] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[9] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, 2000.

[10] D. Wu, Y. T. Hou, B. Li, W. Zhu, Y.-Q. Zhang, and H. J. Chao, "An end-to-end approach for optimal mode selection in internet video communication: theory and application," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 977–995, 2000.

[11] S. Mao, Y. T. Hou, X. Cheng, H. D. Sherali, S. F. Midkiff, and Y.-Q. Zhang, "On routing for multiple description video over wireless ad hoc networks," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 1063–1074, 2006.

[12] A. C. Begen, Y. Altunbasak, O. Ergun, and M. H. Ammar, "Multi-path selection for multiple description video streaming over overlay networks," *Signal Processing: Image Communication*, vol. 20, no. 1, pp. 39–60, 2005.

[13] Y. Andreopoulos, N. Mastronarde, and M. van der Schaar, "Cross-layer optimized video streaming over wireless multihop mesh networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2104–2115, 2006.

[14] J. Villalón, P. Cuenca, L. Orozco-Barbosa, Y. Seok, and T. Turletti, "Cross-layer architecture for adaptive video multicast streaming over multirate wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 699–711, 2007.

[15] J. Wang, M. Venkatachalam, and Y. Fang, "System architecture and cross-layer optimization of video broadcast over WiMAX," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 712–721, 2007.

[16] Z. Li, G. M. Schuster, A. K. Katsaggelos, and B. Gandhi, "Rate-distortion optimal video summary generation," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1550–1560, 2005.

[17] P. V. Pahalawatta, Z. Li, F. Zhai, and A. K. Katsaggelos, "Rate-distortion optimized video summary generation and transmission over packet lossy networks," in *Image and Video Communications and Processing (IVCP '05)*, vol. 5685 of *Proceedings of SPIE*, pp. 801–809, San Jose, Calif, USA, January 2005.

[18] M. van der Schaar, D. S. Turaga, and R. Wong, "Classification-based system for cross-layer optimized wireless video transmission," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 1082–1095, 2006.

[19] A. Luthra, G. J. Sullivan, and T. Wiegand, "Introduction to the special issue on the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 557–559, 2003.

[20] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[21] Q. Xu, V. Stanković, and Z. Xiong, "Distributed joint source-channel coding of video using raptor codes," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 851–861, 2007.

[22] S. Choudhury and J. D. Gibson, "Payload length and rate adaptation for multimedia communications in wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 796–807, 2007.

[23] N. Mastronarde, D. S. Turaga, and M. van der Schaar, "Collaborative resource exchanges for peer-to-peer video streaming over wireless mesh networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 1, pp. 108–118, 2007.

[24] J. Chen and Y. Yang, "Multi-hop delay performance in wireless mesh networks," in *Proceedings of IEEE International Conference on Communications (ICC '07)*, Glasgow, Scotland, June 2007.

[25] C. Wang, B. Li, K. Sohraby, M. Daneshmand, and Y. Hu, "Upstream congestion control in wireless sensor networks through cross-layer optimization," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 786–795, 2007.

[26] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, 2003.

[27] S. Kompella, S. Mao, Y. T. Hou, and H. D. Sherali, "Cross-layer optimized multipath routing for video communications in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 831–840, 2007.

[28] B. Liang and M. Dong, "Packet prioritization in multihop latency aware scheduling for delay constrained communication," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 819–830, 2007.

[29] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1142–1153, 2005.

[30] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Content-aware resource allocation and packet scheduling for video transmission over wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 749–759, 2007.

[31] H.-P. Shiang and M. van der Schaar, "Multi-user video streaming over multi-hop wireless networks: a distributed, cross-layer approach based on priority queuing," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 770–785, 2007.

[32] S. Mohapatra, N. Dutt, A. Nicolau, and N. Venkatasubramanian, "DYNAMO: a cross-layer framework for end-to-end QoS and energy optimization in mobile handheld devices," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 722–737, 2007.

[33] D. Rajan, "Towards universal power efficient scheduling in Gaussian channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 808–818, 2007.

[34] Y.-J. Chang, F.-T. Chien, and C.-C. J. Kuo, "Cross-layer QoS analysis of opportunistic OFDM-TDMA and OFDMA networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 657–666, 2007.

[35] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1746–1755, 2004.

[36] P. Werbose, "ADP: goals, oppotunities and principles," in *Handbook of Learning and Approaximate Dynamic Programming*, J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, Eds., Wiley-IEEE Press, New York, NY, USA, 2004.

[37] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, USA, 1957.

[38] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, Mass, USA, 1996.

[39] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, Eds., *Handbook of Learning and Approaximate Dynamic Programming*, Wiley-IEEE Press, New York, NY, USA, 2004.

[40] D. Denneberg, *Non-Additive Measure and Integral*, Kluwer Academic, Dordrecht, The Netherlands, 1994.

[41] D. Denneberg and M. Grabisch, "Interaction transform of set functions over a finite set," *Information Sciences*, vol. 121, no. 1-2, pp. 149–170, 1999.

[42] M. Grabisch, "Fuzzy measures and integrals: a survey of applications and recent issues," in *Fuzzy Sets Methods in Information Engineering: A Guided Tour of Applications*, D. Dubois, H. Prade, and R. Yager, Eds., John Wiley & Sons, New York, NY, USA, 1996.

[43] R. R. Yager, "Uncertainty representation using fuzzy measures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 32, no. 1, pp. 13–20, 2002.

[44] G. Choquet, "Theory of capacities," *Annales de l'Institut Fourier*, vol. 5, pp. 131–295, 1953.

[45] J.-R. Chang, C.-T. Hung, and G.-H. Tzeng, "Non-additive grey relational model: case study on evaluation of flexible pavement," in *Proceedings of IEEE International Conference on Fuzzy Systems*, vol. 1, pp. 577–582, Budapest, Hungary, July 2004.

[46] A. Denguir-Rekik, G. Mauris, and J. Montmain, "Propagation of uncertainty by the possibility theory in Choquet integral-based decision making: application to an E-commerce website choice support," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 3, pp. 721–728, 2006.

[47] M.-H. Ha, Z.-F. Feng, E.-L. Du, and Y.-C. Bai, "Further discussion on quasi-probability," in *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC '06)*, pp. 3508–3513, Dalian, China, August 2006.

[48] Z. Wang, "A new model of nonlinear multiregressions by projection pursuit based on generalized Choquet integrals," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ '02)*, vol. 2, pp. 1240–1244, Honolulu, Hawaii, USA, May 2002.

[49] M. Grabisch, "k-order additive discrete fuzzy measures and their representation," *Fuzzy Sets and Systems*, vol. 92, no. 2, pp. 167–189, 1997.

[50] M. Grabisch, "Modelling data by the Choquet integral," in *Information Fusion in Data Mining*, V. Torra, Ed., pp. 135–148, Physica, Heidelberg, Germany, 2003.

[51] D. Wu, S. Ci, and H. Wang, "Cross-layer optimization for video summary transmission over wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 841–850, 2007.

[52] IETF, "RFC 3626—optimized link state routing protocol (OLSR)," http://www.faqs.org/rfcs/rfc3626.html.

[53] S. Ci and H.-F. Guo, "Quantitative dynamic interdependency measure and significance analysis for cross-layer design under uncertainty," in *Proceedings of the 16th International Conference on Computer Communications and Networks (ICCCN '07)*, pp. 900–904, Honolulu, Hawaii, USA, August 2007.

[54] M. Grabisch, "A graphical interpretation of the Choquet integral," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 5, pp. 627–631, 2000.

[55] S. Ci and H.-F Guo, "Significance measure with nonlinear and incommensurable observations," in *Proceedings of IEEE Global Communications Conference (GLOBECOM '07)*, Washington, DC, USA, November 2007.