Research Trends in Multimedia Systems and Hardware Accelerators for AI

Nabil J. Sarhan
The Department of Electrical and Computer Engineering
Wayne State University
Detroit, MI, 48202, USA
e-mail: nabil.sarhan@wayne.edu

September 2024

1 Introduction

This paper summarizes the main contributions of the Wayne State University Computer Systems and Deep Learning Lab in the areas of the design of multimedia systems [35, 32, 29, 29, 6, 41, 49, 34], the design of computer vision systems [16, 15], and hardware accelerators for AI [25, 20, 19, 24, 18, 23, 57, 59, 58, 60, 36, 26].

2 Design of Multimedia Systems

The main work on multimedia systems can be summarized as follows: and video coding [51, 14, 54], supporting heterogeneous receivers [35, 32], client-side caching [28], supporting interactive operations [29, 27], supporting advertisements [4, 2, 40, 1], waiting time prediction [6, 8, 41, 9, 40], request scheduling [49, 7, 45, 39], proving time-of-service guarantees [47, 44, 48], resource sharing and server-side cache management [34, 33, 31], streaming media workload characterization [30], and storage subsystem support [38, 46, 43, 42] . The main principles of multimedia streaming are discussed in [37].

We discuss next some of our major work on the design of computer systems for multimedia applications.

2.1 Video Coding

Study [54] describes the major differences between VP8 and H.264 and provides detailed comparative evaluations through extensive experiments. We use 29 raw video sequences, offering a wide spectrum of resolutions and content characteristics, with resolution ranging from 176x144 (QCIF) to 3840x2160 (2160p). To ensure a fair study, we use 3 coding presets in H.264, each with three types of tuning, and 7 presets in VP8. The presets cover a variety of achieved quality or complexity levels. The performance metrics include accuracy of bitrate handling, encoding speed, decoding speed, and perceptual video quality.

In study [51], we introduce a state-of-the-art adaptive instant learning-based model, named iHELP, developed to address the computational complexity arising from encoders' adaptive block structures. The iHELP model achieves outstanding coding efficiency and quality while considerably improving encoding speed. iHELP model has been tested on High Efficiency Video Codec

(HEVC), but it applies to other encoders with similar adaptive block structures. iHELP model employs entropy-based block similarity to predict the splitting decision of the Large Coding Units (LCU), determining whether to divide the block based on the correlation between the block content and previously adjacent encoded blocks in both spatial and temporal dimensions. Our methodology has been rigorously evaluated using the HEVC standard's common test conditions, and the results indicate that iHELP serves as an effective solution for efficient video coding in bandwidth-constrained situations, making it suitable for real-time video applications. The proposed method achieves an 80% reduction in encoding time while maintaining comparable Peak Signal-to-Noise Ratio (PSNR) relative to the Rate-Distortion Optimization (RDO) approach. The exceptional potential of the iHELP model calls for further exploration, as no other existing methods have demonstrated such a high level of performance.

Study [14] proposes a deep learning based methodology to enhance the availability of video streaming systems by developing a prediction model for video streaming quality, required power consumption, and required bandwidth based on video codec parameters. H.264/AVC codec is chosen as a case study. We model the predicted consumed power, the predicted perceived video quality, and the predicted required bandwidth for the video codec based on video resolution and quantization parameters. We train, validate, and test the developed models through extensive experiments using several video contents. Results show that an accurate model can be built for the needed purpose and the video streaming quality, required power consumption, and required bandwidth can be predicted accurately which can be utilized to enhance network availability in a cooperative environment.

2.2 Supporting Heterogeneous Receivers

The number of video streams that can be serviced concurrently is highly constrained by the required real-time and high-rate transfers of multimedia data. Resource sharing techniques, such as Batching, Patching, and Earliest Reachable Merge Target (ERMT), can be used to address this problem by utilizing the multicast facility, which allows multiple requests to share the same set of server and network resources. They assume, however, that all clients have the same available download bandwidth and buffer space. In [35, 32], we study how to efficiently support clients with varying available download bandwidth and buffer space, while delivering data in a client-pull fashion using enhanced resource sharing. In particular, we propose three hybrid solutions to address the variability in the download bandwidth among clients: Simple Hybrid Solution (SHS), Adaptive Hybrid Solution (AHS), and Enhanced Hybrid Solution (EHS). SHS simply combines Batching with either Patching or ERMT, leading to two alternatives: SHS-P and SHS-E, respectively. Batching is used for clients with bandwidth lower than double the video playback rate, and Patching/ERMT is used for the rest. In contrast, AHS and EHS classify clients into multiple bandwidth classes and service them accordingly. AHS employs a new stream type, called adaptive stream, and EHS employs an enhanced adaptive stream type to serve clients with bandwidth capacities ranging between the video playback rate and double that rate. AHS and EHS employ adaptive streams or enhanced adaptive streams in conjunction with Batching and Patching or ERMT, leading to four possible schemes: AHS-P, AHS-E, EHS-P, and EHS-E. Moreover, we consider the variability of the available buffer space among clients. Furthermore, we study how the waiting playback requests for different videos can be scheduled for service in the heterogeneous environment, capturing the variations in both the client bandwidth and buffer space. We evaluate the effectiveness of the proposed solutions and analyze various scheduling policies through extensive simulation.

2.3 Waiting-Time Prediction

Providing video streaming users with expected waiting times enhances their perceived quality-of-service (QoS) and encourages them to wait. In the absence of any waiting-time feedback, users are more likely to defect because of the uncertainty as to when their services will start.

In [6, 41], we analyze waiting-time predictability in scalable video streaming. We propose two prediction schemes and study their effectiveness when applied with various stream merging techniques and scheduling policies. The results demonstrate that the waiting time can be predicted accurately, especially when enhanced cost-based scheduling is applied. The combination of waiting-time prediction and cost-based scheduling leads to outstanding performance benefits

2.4 Client-Side Caching

The design of interactive Near Video-on-Demand (NVOD) systems is highly complicated when scalable stream merging is used. We propose an intelligent client-side cache management policy for these systems, allowing and exploiting cache discontinuity. This policy maximizes the percentage of interactive requests serviced from the client's own cache without requiring any resources from the server. The policy caches data from all streams that are being listened to by the client. As the cache becomes full, it purges data according to a purging algorithm. In [28], we present three purging algorithms: purge the oldest data, purge the furthest data from the customer's playback point, and purge adaptively. Moreover, we experiment with another important decision, which is whether pausing users should continue to listen to streams when the cache becomes full. We evaluate the effectiveness of the proposed cache management policy and purging algorithms under realistic and complex workload through extensive simulations. We analyze many metrics, including waiting and blocking metrics, aggregate delay, cache hit rate, and cache fragment

2.5 Supporting Interactive Operations

The required real-time and high-rate transfer of multimedia data limits the number of requests that can be concurrently serviced by video-on-demand (VOD) systems. Resource-sharing techniques can be used to address this scalability challenge, but they greatly complicate the efficient support for interactive operations. In [29, 27], we develop an overall solution for interactive near VOD systems that employ resource sharing. The proposed solution supports user interactions with short response times and low rejection probabilities. The solution includes a novel stream provisioning policy, which dynamically determines the best number of I-Streams (unicast streams for supporting interactive requests) and the maximum I-Stream length that can be allocated by the server. Furthermore, we use a sophisticated client-side cache management policy to maximize the percentage of interactive requests serviced from the client's own cache. We study the system using a realistic workload through extensive simulation.

2.6 Supporting Advertisements

Studies [4, 2, 40, 1] develop a scalable delivery solution for commercial near-on-demand video streaming systems with an associated pricing model. The proposed delivery solution combines the benefits of periodic broadcasting and stream merging, thereby enabling scalable video delivery. Video advertisements are delivered to the clients prior to viewing the requested videos. The revenues generated from the ads are used to subsidize the price of the requested videos. The pricing is determined based on the total ad viewing time. The proposed solution includes an efficient

ad allocation scheme and a new constraint-based scheduling approach. In addition, they investigate how targeted advertisements can be efficiently supported. Furthermore, they investigate the effectiveness of the overall solutions and analyze and compare the effectiveness of various scheduling policies and ad allocation alternatives in terms of several metrics, including client defection probability, average number of viewed ads per client, price, channel utilization, revenue, and profit.

2.7 Request Scheduling

The number of media streams that can be supported concurrently is highly constrained by the stringent requirements of real-time playback and high transfer rates. To address this problem, media delivery techniques, such as Batching and Stream Merging, utilize the multicast facility to increase resource sharing. The achieved resource sharing depends greatly on how the waiting requests are scheduled for service. Scheduling has been studied extensively when Batching is applied, but up to our knowledge, it has not been investigated in the context of stream merging techniques, which achieve much better resource sharing. In [49], we analyze scheduling when stream merging is employed and propose a simple, yet highly effective scheduling policy, called Minimum Cost First (MCF). MCF exploits the wide variation in stream lengths by favoring the requests that require the least cost. We present two alternative implementations of MCF: MCF-T and MCF-P. We compare various scheduling policies through extensive simulation and show that MCF achieves significant performance benefits in terms of both the number of requests that can be serviced concurrently and the average waiting time for service.

2.8 Resource Sharing and Server-Side Cache Management

The required real-time and high-rate transfers for multimedia data severely limit the number of video streams that can be delivered concurrently. Resource-sharing techniques address this problem and can be classified into two main classes: stream merging and periodic broadcasting. In [34], we evaluate major resource-sharing techniques from the two classes, considering different service models and video workloads. We utilize this extensive analysis in developing a workload-aware hybrid solution (WAHS) that combines the advantages of the best performers among resource-sharing techniques. Moreover, we propose a statistical cache management (SCM) approach and derive analytical models for optimal cache allocation to reduce further the demands on the disk I/O when various resource sharing techniques are used.

2.9 Providing Time-of-Service Grantees

Recent advances in storage and communication technologies have spurred a strong interest in Video-on-Demand (VOD) services. Providing the customers of VOD servers with time of service guarantees offers two major advantages. First, it makes VOD services more attractive by improving customer-perceived quality of service (QoS). Second, it improves throughput through the enhanced resource sharing attained by motivating the customers to wait.

In [47, 44, 48], we propose a new class of scheduling policies, called *Next Schedule Time First (NSTF)*, which provides customers with schedule times and performs scheduling based on these schedule times. NSTF guarantees that customers will be serviced no later than scheduled and ensures that the schedule times are very accurate estimates of the actual times of service. We present alternative implementations of NSTF and show through simulation that NSTF works as expected and delivers outstanding performance benefits.

2.10 Streaming Media Workload Characterization

The popularity of social media has grown dramatically over the World Wide Web. In [30], we analyze the video popularity distribution of well-known social video websites and characterize their workload. We identify trends in the categories, lengths, and formats of those videos, as well as characterize the evolution of those videos over time. We further provide an extensive analysis and comparison of video content amongst the main regions of the world.

2.11 Storage Subsystem Support

Video streaming servers waste precious resources in performing store-and-forward copying. This excessive overhead increases cost and severely limits the scalability of these servers. In [46, 43], we propose using the Network-Attached Disk (NAD) architecture to design highly scalable and cost-effective video streaming servers. To ensure enhanced performance, we propose a scheme, called Distributed Interval Caching (DIC), which utilizes the on-disk buffers for caching intervals between successive streams. We also propose another scheme, called Multi-Objective Scheduling (MOS), which increases the degrees of resource sharing by scheduling the waiting requests for service intelligently. We then integrate the two schemes and study the overall performance benefits through extensive simulation. The results demonstrate that the integrated policy works very well in increasing the number of customers that can be serviced concurrently while decreasing their waiting times for service. The performance benefits vary with several architectural, system workload, and scheduling parameters. We conclude this study by developing an analytical model for ideal DIC to estimate the performance limits which may be achieved through various optimizations.

In [42], we exploit video access patterns and propose an adaptive rearrangement of the blocks on each disk within the server. With this approach, the blocks of the videos with comparable access frequencies are kept closer to each other. We analyze two rearrangement schemes: centered-layout and sequential layout. In the centered layout, blocks are placed according to their access patterns starting with the most popular movie at the center. The sequential layout places movies in the order of their popularity starting at the edge of the disk.

2.12 Power Consumption Modeling

Study [55] analyzes and compares the rate-accuracy and rate-energy characteristics of various video rate adaptation techniques in computer vision applications. The analyzed rate adaptation techniques include spatial, spatial with upscaling, temporal, and Signal-to-Noise Ratio (SNR). We experiment with standard video sequences as well as 300 security, surveillance, news, and speech videos. These videos total 19.15 hours of recording time. We consider both MPEG-4 and H.264 compression standards.

Study [50] models energy consumption of H.264/AVC and VP8 encoders utilizing machine learning.

3 Design of Computer Vision Systems

The main work on computer vision systems, such as automated video surveillance systems, includes the following: video rate adaptation [53, 52, 22], cross-layer optimization [11, 10, 17, 12, 13, 16], autonomous control of PTZ cameras for optimal threat detection accuracy [3, 5, 15], characterization of deep learning accuracy [22, 21], and power consumption modeling [55, 56],

We discuss next some of our major work on the design of computer vision systems.

3.1 Video Rate Adaptation

Video rate adaptation is analyzed in [53, 52, 22]. These papers analyze and compare the rate-accuracy and rate-energy characteristics of various video rate adaptation techniques in computer vision applications. The analyzed rate adaptation techniques include spatial, spatial with upscaling, temporal, and Signal-to-Noise Ratio (SNR). We experiment with standard video sequences as well as security, surveillance, news, and speech videos. These videos total 19.15 hours of recording time. We consider both MPEG-4 and H.264 compression standards.

3.2 Cross-layer Optimization

Studies [10, 12] develop a cross-layer optimization framework for video streaming from multiple sources to a central proxy station over a wireless network. The proposed framework manages the application rates and transmission opportunities of various video sources based on the dynamic network conditions in such a way that minimizes the overall distortion. The framework utilizes a novel online approach for estimating the effective airtime of the network. We demonstrate the effectiveness of the proposed framework and effective airtime estimation approach through extensive experiments.

Studies [11, 13] develop an accuracy-based cross-layer optimization solution for wireless automated video surveillance systems, in which multiple sources stream videos to a central proxy station. The proposed solution manages the application rates and transmission opportunities of various video sources based on the dynamic network conditions in such a way that maximizes the overall detection accuracy of the computer vision algorithm(s). We demonstrate the effectiveness of the proposed solution through extensive simulations.

Studies [17, 16] consider video analytics systems in which a central monitoring station receives and analyzes the video streams captured and delivered wirelessly by multiple cameras. They address how the bandwidth can be allocated to various cameras by presenting a cross-layer solution that optimizes the overall detection or recognition accuracy. In further contrast with prior work, they present and develop a real CV system and subsequently provide a detailed experimental analysis of cross-layer optimization. Other unique features of the developed solution include employing the popular HTTP streaming approach, utilizing homogeneous cameras as well as heterogeneous ones with varying capabilities and limitations, and including a new algorithm for estimating the effective medium airtime. The results show that the proposed solution significantly improves the CV accuracy.

3.3 Autonomous Control of PTZ Cameras for Optimal Threat Detection

Studies [3, 15] address the research problem of how to autonomously control Pan/Tilt/Zoom (PTZ) cameras in a manner that seeks to optimize the face recognition accuracy or the overall threat detection and proposes an overall system. The papers present two alternative schemes for camera scheduling: Grid-Based Grouping (GBG) and Elevator-Based Planning (EBP). The camera control works with realistic 3D environments and considers many factors, including the direction of the subject's movement and its location, distances from the cameras, occlusion, overall recognition probability so far, and the expected time to leave the site, as well as the movements of cameras and their capabilities and limitations.

Study [15] utilizes clustering to group subjects, thereby enabling the system to focus on the areas that are more densely populated. The clustering approach is detailed in [5]. Moreover, it proposes a dynamic mechanism for controlling the pre-recording time spent on running the solution.

Furthermore, it develop a parallel algorithm, allowing the most time-consuming phases to be parallelized and thus run efficiently by the centralized parallel processing subsystem. We analyze through simulation the effectiveness of the overall solution, including the clustering approach, scheduling alternatives, dynamic mechanism, and parallel implementation in terms of overall recognition probability and the running time of the solution, considering the impacts of numerous parameters.

3.4 Analytical Modeling of Deep Learning Accuracy in Adaptive Video Streams

To fit the tight resource constraints, including network bandwidth, the video streams in Computer Vision systems are adapted dynamically by changing the video capturing and encoding parameters. In [22, 21], we propose two novel analytical models that characterize the face recognition accuracy in terms of these parameters, specifically resolution, quantization, and actual bitrate. We find that the accuracy is a logistic function of the video quantization parameter, with the value of the Sigmoid's midpoint being a function of the resolution. Alternatively, we find that the accuracy is equal to the sum of two exponentials of the actual video bitrate, with the resolution as a multiplicative factor with one exponential. We develop an evaluation framework to validate the models using two distinct video datasets with 99 videos and the widely used Labeled Faces in the Wild (LFW) dataset with $13,233$ images. We conduct $1,668$ experiments that involve varying combinations of encoding parameters. We show that both models hold true for the deep-learning and statistical-based face recognition. The developed models achieve an average coefficient of determination (R2) of 98.7% to 99.8%.

Study [56] develops an aggregate power consumption model for live video streaming systems, including many-to-many systems. In many-to-one streaming systems, multiple video sources (i.e., cameras and/or sensors) stream videos to a monitoring station. We model the power consumed by the video sources in the capturing, encoding, and transmission phases and then provide an overall model in terms of the main capturing and encoding parameters, including resolution, frame rate, number of reference frames, motion estimation range, and quantization. We also analyze the power consumed by the monitoring station due to receiving, decoding, and upscaling the received video streams. In addition to modeling the power consumption, we model the achieved bitrate of video encoding. We validate the developed models through extensive experiments using two types of systems and different video contents. Furthermore, we analyze many-to-one systems in terms of bitrate, video quality, and the power consumed by the sources, as well as that by the monitoring station, considering the impacts of multiple parameters simultaneously.

4 Hardware Accelerators for AI

The research lab has multiple contributions in other research areas, including the design of hardware accelerators for AI [25, 20, 19, 24, 18, 23] and epileptic seizure prediction through deep learning [57, 59, 58, 60, 36, 26].

References

[1] Musab Al-Hadrusi and Nabil J. Sarhan. Scalable delivery and pricing of streaming media with advertisements. In *Proceedings of the 15th International Conference on Multimedia (ACM MM 2007)*, pages 791–794. ACM, September 2007.

[2] Musab Al-Hadrusi and Nabil J. Sarhan. Client-driven price selection for scalable video streaming with advertisements. In *Proceedings of the 18th International MultiMedia Modeling Con-*

ference (MMM 2012), volume 7131 of *Lecture Notes in Computer Science*, pages 429–439. Springer, January 2012.

[3] Musab Al-Hadrusi and Nabil J. Sarhan. Efficient control of PTZ cameras in automated video surveillance systems. In *Proceedings of the 2012 IEEE International Symposium on Multimedia (ISM 2012)*, pages 356–359. IEEE Computer Society, December 2012.

[4] Musab Al-Hadrusi and Nabil J. Sarhan. A scalable delivery solution and a pricing model for commercial video-on-demand systems with video advertisements. *Multimedia Tools and Applications*, 73(3):1417–1443, 2014.

[5] Musab S. Al-Hadrusi, Nabil J. Sarhan, and Sina G. Davani. A clustering approach for controlling PTZ cameras in automated video surveillance. In *Proceedings of the IEEE International Symposium on Multimedia (ISM 2016)*, pages 333–336. IEEE Computer Society, December 2016.

[6] Mohammad A. Alsmirat, Musab Al-Hadrusi, and Nabil J. Sarhan. Analysis of waiting-time predictability in scalable media streaming. In *Proceedings of the 15th International Conference on Multimedia (ACM MM 2007)*, pages 727–736. ACM, September 2007.

[7] Mohammad A. Alsmirat and Nabil J. Sarhan. Predictive cost-based scheduling for scalable media streaming. In *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo (ICME 2008)*, pages 857–860. IEEE Computer Society, June 2008.

[8] Mohammad A. Alsmirat and Nabil J. Sarhan. Performance and waiting-time predictability analysis of design options in cost-based scheduling for scalable media streaming. In *Proceedings of the 15th International Multimedia Modeling Conference (MMM 2009)*, volume 5371 of *Lecture Notes in Computer Science*, pages 150–162. Springer, January 2009.

[9] Mohammad A. Alsmirat and Nabil J. Sarhan. Detailed performance and waiting-time predictability analysis of scheduling options in on-demand video streaming. *EURASIP Journal on Image and Video Processing*, 2010.

[10] Mohammad A. Alsmirat and Nabil J. Sarhan. Cross-layer optimization and effective airtime estimation for wireless video streaming. In *Proceedings of the 21st International Conference on Computer Communications and Networks (ICCCN 2012)*, pages 1–7. IEEE, July- August 2012.

[11] Mohammad A. Alsmirat and Nabil J. Sarhan. Cross-layer optimization for automated video surveillance. In *Proceedings of the IEEE International Symposium on Multimedia (ISM 2016)*, pages 243–246. IEEE Computer Society, December 2016.

[12] Mohammad A. Alsmirat and Nabil J. Sarhan. Cross-layer optimization for many-to-one wireless video streaming systems. *Multimedia Tools and Applications*, 77(19):24789–24811, 2018.

[13] Mohammad A. Alsmirat and Nabil J. Sarhan. Intelligent optimization for automated video surveillance at the edge: A cross-layer approach. *Simulation Modelling Practice and Theory*, 105:102171, 2020.

[14] Mohammad A. Alsmirat, Yousef O. Sharrab, Monther Tarawneh, Sana'a Al-shboul, and Nabil J. Sarhan. Video coding deep learning-based modeling for long life video streaming over next network generation. *Cluster Computing*, 26(2):1159–1167, 2023.

[15] Sina G. Davani, Musab S. Al-Hadrusi, and Nabil J. Sarhan. An autonomous system for efficient control of PTZ cameras. *ACM Transactions on Autonomous and Adaptive Systems*, 16(2):6:1–6:22, 2021.

[16] Sina G. Davani and Nabil J. Sarhan. Experimental analysis of optimal bandwidth allocation in computer vision systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4121–4130, 2021.

[17] Sina Gholamnejad Davani and Nabil J. Sarhan. Experimental analysis of bandwidth allocation in automated video surveillance systems. In *Proceedings of the 2017 ACM on Multimedia Conference (ACM MM 2017)*, pages 1457–1464. ACM, October 2017.

[18] Melvin D. Edwards, Hamza Al Maharmeh, Nabil J. Sarhan, Mohammed Ismail, and Mohammad Alhawari. A low-power, digitally-controlled, multi-stable, CMOS analog memory circuit. In *Proceedings of the 63rd IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2020)*, pages 872–875. IEEE, August 2020.

[19] Melvin D. Edwards, Nabil J. Sarhan, and Mohammad Alhawari. Analysis of dual-row and dual-array crossbars in mixed signal deep neural networks. In *Proceedings of the 66th IEEE International Midwest Symposium on Circuits and Systems, MWSCAS 2023, Tempe, AZ, USA, August 6-9, 2023*, pages 654–658. IEEE, 2023.

[20] Melvin D. Edwards, Nabil J. Sarhan, and Mohammad Alhawari. A CMOS analog neuron circuit with A multi-level memory. In *International Conference on Microelectronics, ICM 2023, Abu Dhabi, United Arab Emirates, December 17-20, 2023*, pages 11–15. IEEE, 2023.

[21] Hayder Hamandi and Nabil J. Sarhan. QRMODA and BRMODA: novel models for face recognition accuracy in computer vision systems with adapted video streams. *CoRR*, abs/1907.10559, 2019.

[22] Hayder R. Hamandi and Nabil J. Sarhan. Novel analytical models of face recognition accuracy in terms of video capturing and encoding parameters. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2020)*, pages 1–6. IEEE, July 2020.

[23] Hamza Al Maharmeh, Nabil J. Sarhan, Chung-Chih Hung, Mohammed Ismail, and Mohammad Alhawari. Compute-in-time for deep neural network accelerators: Challenges and prospects. In *Proceedings of the 63rd IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2020)*, pages 990–993. IEEE, August 2020.

[24] Hamza Al Maharmeh, Nabil J. Sarhan, Chung-Chih Hung, Mohammed Ismail, and Mohammad Alhawari. A comparative analysis of time-domain and digital-domain hardware accelerators for neural networks. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2021)*, pages 1–5. IEEE, May 2021.

[25] Hamza Al Maharmeh, Nabil J. Sarhan, Mohammed Ismail, and Mohammad Alhawari. A 116 TOPS/W spatially unrolled time-domain accelerator utilizing laddered-inverter DTC for energy-efficient edge computing in 65 nm. *IEEE Open J. Circuits Syst.*, 4:308–323, 2023.

[26] Ian McNulty, Shiva Maleki Varnosfaderani, Omar Makke, Nabil J. Sarhan, Eishi Asano, Aimee F. Luat, and Mohammad Alhawari. Analysis of artifacts removal techniques in EEG signals for energy-constrained devices. In *Proceedings of the 64th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2021)*, pages 515–519. IEEE, August 2021.

[27] Kamal K. Nayfeh and Nabil J. Sarhan. Design and analysis of scalable and interactive near video-on-demand systems. In *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME 2013)*, pages 1–6. IEEE Computer Society, July 2013.

[28] Kamal K. Nayfeh and Nabil J. Sarhan. Client-side cache management for scalable and interactive video streaming. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2016)*, pages 1–6. IEEE Computer Society, July 2016.

[29] Kamal K. Nayfeh and Nabil J. Sarhan. A scalable solution for interactive near video-on-demand systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(10):1907–1916, 2016.

[30] Jeffrey R. Ostrowski and Nabil J. Sarhan. Characterization of social video. In *Proceedings of SPIE Multimedia Computing and Networking Conference (MMCN)*, volume 7253, page 72530E. International Society for Optics and Photonics, SPIE, January 2009.

[31] Bashar Qudah and Nabil J. Sarhan. Analysis of resource sharing and cache management in scalable video-on-demand. In *Proceedings of the 14th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2006)*, pages 327–334. IEEE Computer Society, September 2006.

[32] Bashar Qudah and Nabil J. Sarhan. Towards scalable delivery of video streams to heterogeneous receivers. In *Proceedings of the 14th ACM International Conference on Multimedia (ACM MM 2006)*, pages 347–356. ACM, October 2006.

[33] Bashar Qudah and Nabil J. Sarhan. Towards enhanced resource sharing in video streaming with generalized access patterns. In *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo (ICME 2007)*, pages 1219–1222. IEEE Computer Society, July 2007.

[34] Bashar Qudah and Nabil J. Sarhan. Workload-aware resource sharing and cache management for scalable video streaming. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(3):386–396, 2009.

[35] Bashar Qudah and Nabil J. Sarhan. Efficient delivery of on-demand video streams to heterogeneous receivers. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(3):20:1–20:25, 2010.

[36] Rihat Rahman, Shiva Maleki Varnosfaderani, Omar Makke, Nabil J. Sarhan, Eishi Asano, Aimee F. Luat, and Mohammad Alhawari. Comprehensive analysis of EEG datasets for epileptic seizure prediction. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2021)*, pages 1–5. IEEE, May 2021.

[37] Nabil J. Sarhan. Multimedia streaming. In *Handbook of Computer Networks*, pages 282–292. John Wiley & Sons, Ltd, 2007.

[38] Nabil J. Sarhan. *On the Design of Scalable and High Performance Multimedia Servers*. Ph.D. Dissertation, Pennsylvania State University, Department of Computer Science and Engineering, August 2003.

[39] Nabil J. Sarhan. An investigation of scheduling policies for multimedia systems. M.S. Thesis, Pennsylvania State University, Department of Computer Science and Engineering, May 2003.

[40] Nabil J. Sarhan and Musab Al-Hadrusi. Waiting-time prediction and QoS-based pricing for video streaming with advertisements. In *Proceedings of the 12th IEEE International Symposium on Multimedia (ISM 2010)*, pages 17–24. IEEE Computer Society, December 2010.

[41] Nabil J. Sarhan, Mohammad A. Alsmirat, and Musab Al-Hadrusi. Waiting-time prediction in scalable on-demand video streaming. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(2):11:1–11:25, 2010.

[42] Nabil J. Sarhan and Chita R. Das. Adaptive block rearrangement algorithms for video-on-demand servers. In *Proceedings of the 2001 International Conference on Parallel Processing (ICPP 2002)*, pages 452–462. IEEE Computer Society, September 2001.

[43] Nabil J. Sarhan and Chita R. Das. An integrated resource sharing policy for multimedia storage servers based on network-attached disks. In *Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS 2003)*, page 136. IEEE Computer Society, May 2003.

[44] Nabil J. Sarhan and Chita R. Das. Providing time of service guarantees in video-on-demand servers. In *Proceedings of the Twelfth International World Wide Web Conference (WWW 2003)*, May 2003.

[45] Nabil J. Sarhan and Chita R. Das. A simulation-based analysis of scheduling policies for multimedia server. In *Proceedings of the 36th Annual Simulation Symposium (ANSS-36 2003)*, pages 183–190. IEEE Computer Society, March - April 2003.

[46] Nabil J. Sarhan and Chita R. Das. Caching and scheduling in NAD-based multimedia servers. *IEEE Transactions on Parallel and Distributed Systems*, 15(10):921–933, 2004.

[47] Nabil J. Sarhan and Chita R. Das. A new class of scheduling policies for providing time of service guarantees in video-on-demand servers. In *Proceedings of the 7th IFIP/IEEE International Conference on Management of Multimedia Networks and Services (MMNS 2004)*, volume 3271 of *Lecture Notes in Computer Science*, pages 127–139. Springer, October 2004.

[48] Nabil J. Sarhan and Chita R. Das. Provision of time of service guarantees in video-on-demand servers. Technical Report, Pennsylvania State University, Department of Computer Science and Engineering, College of Engineering, July 2003.

[49] Nabil J Sarhan and Bashar Qudah. Efficient cost-based scheduling for scalable media streaming. In *Proceedings of the Multimedia Computing and Networking Conference (MMCN 2007)*, volume 6504, pages 123–134. SPIE, January-February 2007.

[50] Yousef O. Sharrab; Mohammad Alsmirat; Bilal Hawashin; Nabil J. Sarhan. Machine learning-based energy consumption modeling andcomparing of h.264 and google vp8 encoders. *International Journal of Electrical and Computer Engineering*, 11(2):1303–1310, 2021.

[51] Mohammad A.; Eljinini Mohammad Ali H.; Sarhan Nabil J. Sharrab, Yousef O.; Alsmirat. ihelp: a model for instant learning of video coding in vr/ar real-time applications. *Multimedia Tools and Applications*, 2024.

[52] Yousef O. Sharrab, Izzat Alsmadi, and Nabil J. Sarhan. Towards the availability of video communication in artificial intelligence-based computer vision systems utilizing a multi-objective function. *Cluster Computing*, 25(1):231–247, 2022.

[53] Yousef O. Sharrab and Nabil J. Sarhan. Accuracy and power consumption tradeoffs in video rate adaptation for computer vision applications. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME 2012)*, pages 410–415. IEEE Computer Society, July 2012.

[54] Yousef O. Sharrab and Nabil J. Sarhan. Detailed comparative analysis of VP8 and H.264. In *Proceedings of the 2012 IEEE International Symposium on Multimedia (ISM 2012)*, pages 133–140. IEEE Computer Society, December 2012.

[55] Yousef O. Sharrab and Nabil J. Sarhan. Aggregate power consumption modeling of live video streaming systems. In *Proceedings of the Multimedia Systems Conference 2013 (ACM MMSys '13)*, pages 60–71. ACM, February - March 2013.

[56] Yousef O. Sharrab and Nabil J. Sarhan. Modeling and analysis of power consumption in live video streaming systems. *ACM Transactions on Multimedia Computing Communications and Applications*, 13(4):54:1–54:25, 2017.

[57] Shiva Maleki Varnosfaderani, Ian McNulty, Nabil J. Sarhan, Waleed Abood, and Mohammad Alhawari. An efficient epilepsy prediction model on european dataset with model evaluation considering seizure types. *IEEE Journal of Biomedical and Health Informatics*, pages 1–13, 2024.

[58] Shiva Maleki Varnosfaderani, Ian McNulty, Nabil J. Sarhan, and Mohammad Alhawari. Artifacts removal techniques for the european ieeg dataset. In *Proceedings of the 66th IEEE International Midwest Symposium on Circuits and Systems, MWSCAS 2023, Tempe, AZ, USA, August 6-9, 2023*, pages 298–302. IEEE, 2023.

[59] Shiva Maleki Varnosfaderani, Rihat Rahman, Nabil J. Sarhan, and Mohammad Alhawari. A self-aware power management model for epileptic seizure systems based on patient-specific daily seizure pattern. In *Proceedings of the International Conference on Microelectronics, ICM 2023, Abu Dhabi, United Arab Emirates, December 17-20, 2023*, pages 91–95. IEEE, 2023.

[60] Shiva Maleki Varnosfaderani, Rihat Rahman, Nabil J. Sarhan, Levin Kuhlmann, Eishi Asano, Aimee F. Luat, and Mohammad Alhawari. A two-layer LSTM deep learning model for epileptic seizure prediction. In *Proceedings of the 3rd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS 2021)*, pages 1–4. IEEE, June 2021.