Multimedia, Mathematics, and Machine Learning II

Rabab Ward (UBC)
Li Deng (Microsoft Research)

5 July - 10 July, 2009

1 Overview of the Workshop and the Field

Following the success of the previous Workshop of Multimedia and mathematics during July 23-28, 2005, the current workshop continues the intensive study of the field and brings together the earlier participants plus additional new prominent researchers, with the expanded theme including machine learning. The expanded theme is to push the state of the art in multimedia processing techniques and multimedia technologies by exploring modern mathematical, pattern recognition, and machine learning methods that have cross-media generality. In particular, we bring prominent researchers as well as tutorial lecturers who have rich working/research experiences in one or more media types and who share the experiences on commonality and differences in the mathematical and machine learning techniques for processing different types of media contents.

Multimedia technologies represent rich applications and interactions among a variety of information sources including audio/music, speech, image/graphics/animation, video, and text/documents/language. They also span over wide ranging information processing tasks including coding/compression, analysis, communication/networking/security, synthesis, user interface, perception/recognition/understanding, and retrieval/mining. Future multimedia technology development will require an increasing level of intelligence, for which mathematical representation, modeling, and learning will play an increasingly important role. This is one of the reasons that we include machine learning, providing a rich set of practical algorithms derived from rigorous mathematical analysis. This forms one principal element in the workshops theme.

The main component of the workshop is a series of presentations and ensuing discussions. Due to the multidisciplinary nature of the subject, we invited two tutorial speakers who are experienced in research with multiple media contents. The remaining presentations are focused on state of the art research in a wide range of subjects on multimedia with related machine learning techniques.

2 Presentation Highlights

2.1 Tutorials

The first tutorial was given by Prof. Bernd Girod of Stanford University, entitled “Mobile Image Matching Recognition Meets Compression.” This is a prime example of integrating recognition and coding tasks that are both common in multimedia research. Specifically, the application is on image retrieval with handheld mobile devices, such as camera phones or PDAs, which are expected to become ubiquitous platforms for visual search and mobile augmented reality applications. For mobile image matching, a visual data base is typically stored at a server in the network. Hence, for a visual comparison, information must be either uploaded from the mobile to the server, or downloaded from the server to the mobile. With relatively slow wireless links, the response time of the system critically depends on how much information must be transferred in both directions. The tutorial reviews recent advances in mobile matching, using a “bag-of-visual-words” approach with robust feature descriptors. The results demonsrate that dramatic speed-ups are possible by considering recognition and compression jointly. Real-time implementations for different example applications are described, such as recognition of landmarks or CD cover, to show the benefit from image processing on the phone, the server, and/or both.

The second tutorial was given by Prof. Hermann Ney of RWTH Aachen University in Germany, entitled "Statistical Methods for image, speech, and language processing: Achievements and open problems." This tutorial gives an overview of the statistical methods underlying the dramatic progress in statistical methods for recognizing image and speech signals and for translating spoken and written languageover the last two decades. In particular, it focuses on the remarkable fact that, for all three tasks, the statistical approach makes use of the same four principles: 1) Bayes decision rule for minimum error rate; 2) probabilistic alignment models, e.g. Hidden Markov models, for handling strings of observations (like acoustic vectors for speech recognition and written words for language translation); 3) training criteria and algorithms for estimating the free model parameters from large amounts of data, and 4) the generation or search process that generates the recognition or translation result. The author points out that most of these methods had originally been designed for speech recognition. However, it has turned out that, with suitable modifications, the same concepts carry over to both language translation and image recognition, which in both cases results in systems with state-of-the-art performance. This tutorial elegantly summarizes the achievements and the open problems in this extremely fertile area of statistical modelling.

2.2 Research-Oriented Presentations

In addition to the two tutorials, there are numerous high-quality presentations at the workshop that are focused on research ideas and applications in various areas of multimedia including image/video, audio/speech, and multimedia security. We now give a summary of these presentations.

Prof. Lina Karam of Arizona State University gave a talk on "Adaptive Rate-Distortion Based Wyner-Ziv Video Coding." Her talk starts with a brief introduction to the area of Distributed Video Coding (DVC), which is also known as Wyner-Ziv Video Coding. Two novel adaptive DVC systems are then presented: a pixel-domain DVC system with a rate-distortion based BitpLAne SelecTive decoding (BLAST-DVC), and a transform-domain DVC system with a rate-distortion based Adaptive QuanTization (AQT-DVC). Coding results and comparisons with existing DVC schemes and with H.264 interframe and intraframe coding are presented to illustrate the performance of the proposed systems.

In the presentation entitled "Rectification-based View Interpolation and Extrapolation for Multiview Video Coding: R-D Analysis and Applications," Prof. Jie Liang of Simon Fraser University applis view interpolation in an emerging area of multiview view coding (MVC). Existing schemes assume all cameras are aligned. These methods do not perform well when neighboring cameras point to different directions. In this talk, the author first derives the theoretical R-D performance of the rectification-based view interpolation. He then applies it to H.264 MVC. To further improve the coding efficiency, he develops a rectification-based view extrapolation for MVC. Finally, he investigates the application of the view interpolation in multiple-description coding of multiview images.

In the presentation entitled "High dimensional consensus," Pro.f Jos M. F. Moura of Carnegie Mellon University considers distributed algorithms that can arise when a large number of agents cooperate to reach a common decision or in sensor networks where a large number of sensors cooperate to process large amounts of collected data. In the last few years there has been intensive research in distributed algorithms for such problems. The presenter describes a general class of distributed algorithms, high dimensional consensus (HDC). He shows how a number of problems of interest including distributed inference (like detection, estimation, or classification,) distributed localization, or several types of consensus algorithms can be cast in the framework of HDC. He discusses the convergence of HDC under a broad set of conditions: deterministic as well as random, as when there is noise in the intersensor communications or links among sensors fail at random times. Finally, he address tradeoffs among network and application parameters and their impact on resource allocation, convergence rate, and topology design.

In the presentation entitled "Rotation-invariant wavelet-based matching of local features, with enhanced tolerance to shifts in location and scale," Prof. Nick Kingsbury of University of Cambridge describesa technique for using dual-tree complex wavelets to obtain rich feature descriptors of keypoints in images. The main aim has been to develop a method for retaining the full phase and amplitude information from the complex wavelet coefficients at each scale, while presenting the feature descriptors in a Fourier-domain form that allows for efficient correlation at arbitrary rotations between the candidate and reference image patches. The feature descriptors are known as Polar-Matching matrices. Recently, he modified the previously proposed approach so that it can be more resilient to errors in keypoint location and scale. These multi-scale feature

descriptors are potentially useful for object detection, recognition, classification and tracking in images and video

Prof. Tsuhan Chen of Cornell University presented “A Graphical-Model Framework for Using Context to Understand Images of People.” The motivation of this research is that when we see other humans, we can quickly make judgements regarding many aspects, including their demographic description and identity if they are familiar to us. We can answer questions related to the activities of, emotional states of, and relationships between people in an image. We draw conclusions based not just on what we see, but also from a lifetime of experience of living and interacting with other people. The presenter proposes contextual features and graphical models for understanding images of people with the objective of providing computers with access to the same contextual information that humans use.

In the presentation of “Genomic/Proteomic Signal Processing for Cancer Classification and Prediction,” Prof. Ray Liu of University of Maryland, College Park discussed an important topic of cancer classification and prediction. DNA microarray and proteomic mass spectrum technologies make it possible to simultaneously monitor thousands of genes/protein expression levels and distribution. A topic of great interest is to study the different expression profiles from cancer patients and normal subjects, by classifying them at gene/protein expression levels. Currently, various clustering methods have been proposed in the literature to classify cancer and normal samples based on microarray data, and they are dominantly data-driven approaches. In this talk, an alternative model-driven approach, named ensemble dependence model, is presented aiming at exploring the group dependence relationship of gene clusters. Because of the limited size of current data, it is not feasible to examine the regulation relationship between all genes. Also, both the microarray gene expression and mass spectrum data are noisy. However, if they are clustered in a right way, the noise level in the resulting cluster expression will be reduced, thus the ensemble dependence dynamics of gene clusters will be revealed. Under the framework of hypothesis-testing, genes dependence relationship as a feature to model is employed to classify cancer and normal samples. The classification scheme is then applied to several real cancer data sets. It is noted that the method yields very promising performance.

A group of presentations at the workshop are focused on speech and audio processing and on the general relationship across various media area, following Prof. Ney’s tutorial.

In the presenation of “From Recognition to Understanding — Expanding the Traditional Scope of Signal Processing,” Dr. Li Deng of Microsoft Research at Redmond first observes that the traditional scope of signal processing as defined in the SPS constitution includes the ”signal” classes of audio, video, speech, image, communication, musical, and ”others”, and includes the ”processing” classes of filtering, coding, transmitting, estimating, detecting, analyzing, recognizing, synthesizing, recording, and reproducing. He argues that in our modern information society, we immerse ourselves with the signal processing techniques and applications that go far beyond the above scope. In the presentation, he constructs a ”matrix” which succinctly represents the traditionally defined scope of signal processing and uses this matrix representation to argue for natural expansion of the signal processing scope. In particular, he advocates the extension of the ”signal” coverage from typical numerical type to symbolic type such as text and documents, and for extension of the ”processing” class from recognition to understanding. He then present a case study which demonstrates principled ways in which the commonly used speech recognition techniques are naturally extended to handle the more challenging problem of speech understanding (with new, problem-specific processing steps added in an integrative manner).

The presentation of “Speech Recognition and Machine Translation: A Comparative Overview” given by Dr. Xiaodong He of Microsoft Research, Redmond summarizes substantial progress made over the last decade in both research and real-world applications of speech recognition and machine translation. Despite conspicuous differences, many problems in speech recognition and in machine translation share a wide range of similarities, and it is of great interests to see techniques in these two fields can be successfully cross-fertilized. In this talk, he discusses the similarities and difference between speech recognition and machine translation. As case studies, three specific technologies that have been successfully applied to both fields are discussed in details: hidden Markov model, template based modeling and decoding, and system combination. Through these examples, he compares the properties of speech and language, and shows how generic sequential pattern recognition technologies could be extended and applied to address the particular needs of speech recognition and machine translation.

In the audio processing area, Prof. George Tzanetakis of University of Victoria gave an entertaining talk on “Computational Ethnomusicology - Expanding the reach of Music Information Retrieval to the musics of

the world." Music Information Retrieval (MIR) is a relatively new research area in multimedia. MIR deals with the analysis and retrieval of music in digital form. It reflects the tremendous recent growth of music-related data digitally available and the consequent need to search within it to retrieve music and musical information efficiently and effectively. Most of existing work in MIR has focused on western classical and popular music as these types of music have the largest commercial interest. In this talk Dr. Tzanetakis describes two case studies in Computational Ethnomusicology that explores how MIR techniques can be applied to the study of non-Western music for which there is no standardized written reference (which is a large percentage of the music of world if not of album sales). The first case study is an automatic analysis of micro-timing in complex Afro-Cuban percussion music using rotation-aware dynamic programming. The second case study is a content and context aware web visualization interface for the study of religious chant. In addition to the technical challenges these projects presented he also discusses the social challenges of Interdisciplinary collaborations between engineering and humanities.

The second interesting talk in the area of audio processing is titled "Machine Hearing - A Research Agenda and an Approach," by Richard F. Lyon of Google Inc. at Mountain View. He points out that the field of machine hearing is still in its infancy, in comparison with the thriving field of machine vision. This unfortunate situation, combined with the availability of good front-end auditory models, provides us the opportunity to make quick progress by leveraging techniques from machine vision to help make progress in research and applications in machine hearing. His project at Google aims to help machine hearing become a first-class academic and commercial field. His group developed applications that will do something useful with all that uninterpretable audio media out there, such as sound tracks of amateur movies. There are three main tactics that help us: (1) Leveraging techniques already developed in the machine-vision and machine-learning fields; (2) Productive interaction with the wider field of hearing research, to keep models honest and motivate better experiments; (3) Focus on applications for which the challenge has to do with what things sound like, as opposed to specialized domain knowledge ("non-speech non-music audio"). He presented some results showing how very-high-dimensionality feature spaces can effectively connect auditory representations to simple but powerful machine learning techniques for a range of applications such as sound ranking from text queries.

The third audio-processing talk, titled "Acoustic Scenes, Complex Modulations, and a New Form of Filtering," is by Prof. Les Atlas of University of Washington. Be it in a restaurant or other reverberant and noisy environment, normal hearing listeners segregate multiple sources, usually strongly overlapping in frequency, well beyond capabilities expected by current computational approaches. What is it that we can learn from this common observation? As is now commonly accepted, the differing dynamical modulation patterns of the sources are key to these powers of separation. But until recently, the theoretical underpinnings for the notion of dynamical modulation patterns have been lacking. He has taken a decades-old and loosely defined concept, called "modulation frequency analysis," and developed a theory which allows for distortion-free separation (filtering) of multiple sound sources with differing dynamics. A key result is that previous assumptions of non-negative and real modulation are not sufficient and, instead, coherent and sparse separation approaches are needed to separate different modulation patterns. These results may have an impact in separation and representation of multiple simultaneous sound streams for speech, audio, hearing loss treatment, and underwater acoustic applications. This research also suggests exciting new and potentially important open theoretical questions for general signal representations, extending beyond acoustic applications and potentially impacting other areas of engineering and physics. The final talk in the area of audio/speech processing is titled "Deep-structured learning in speech processing" by Dr. Dong Yu of Microsoft Research, Redmond. In this talk he reports recent investigations on ways to learn complexy sequential decision boundaries in speech processing by composing multiple-layers of simple learners. He shows that this hierarchical structure allows one to learn and use long-range dependencies hidden in the signals and use features that cannot be easily incorporated in the hidden Markov model.

There are a large number of image or video processing related presentations. The first one is "Material classification using visible and near-infrared images," by Prof. Sabine Ssstrunk of EPFL. Recently, the presenter's research group have shown the advantages of simultaneously capturing visible and near infrared (NIR) radiation in digital photography applications, such as white balancing, shadow detection, dehazing, and face rendering. In this talk, she presented the on-going research using NIR images in conjunction with visible images for material classification. As many colorants are transparent to NIR, it is possible to reproduce the intrinsic lightness and texture characteristics of a material. The researchers are thus currently analyzing visible and NIR images according to their lightness and texture. The results are the input of a classifier in the

form of feature vectors, and the probability of that data belong to a material category is then calculated. They achieve good classification results on a limited set of material classes.

The next presentation is "Visualizing and Understanding Challenges in the Design of Light Field Displays," by Dr. Amir Said of Hewlett Packard. While attempts to recreate three-dimensional views are nearly as old as photography, no solution has been able to generate consistent interest and wide acceptance of their quality. New analysis techniques are presented that can easily and more naturally show why the problem is not impossible, but can be indeed very challenging. Sequences of display simulations are shown, which can provide a much more intuitive appreciation of the difficulties, and facilitate understanding how design limitations impact visual quality.

Dr. Jeffrey Bloom of Dialogic Research Inc. presented the next talk of "Understudied Constraints Imposed by Watermarking Applications." As video watermarking becomes more mature and more widely known and accepted, a number of security and non-security applications are emerging. When watermarked content is traveling through a network or series of networks, there is a need to embed and/or detect watermarks at various points in the distribution chain. Traditional watermarking research concentrated only on the input and output of the network. This leaves a number of scenarios understudied. We will discuss two such scenarios: embedding in an entropy-encoded bitstream and detection in a compressed domain that differs from the embedding domain due to transcoding.

The final series of presentations are on multimedia security, starting with Prof. Edward J. Delp (Purdue University) talk on "Multimedia Security: A Viewpoint from a Walking Wounded." This talk describes current research issues in multimedia security involving data hiding, device forensics, biometrics, DRM, and authentication. He "predicts" the future as to where this is all going. This talk also presents a brief overview of the research done in the Video and Image Processing Laboratory at Purdue University. Projects described include video compression, media indexing, multimedia security, language translation, mobile applications, and medical imaging.

The next talk in this series is "Information Management and Security in Media-Sharing Social Networks" by Professors Mehrdad Fatourechi and Jan Wang of University of British Colombia and Prof. Hong Zhao of University of Alberta. Digital media has profoundly changed our daily life during the last decade. For example, the wide adoption of broadband residential access and recent advances in video compression technologies has fueled increasing popularity in delivery of TV services via Internet. We have also witnessed the emergence of large-scale multimedia social network communities such as Facebook and YouTube. This proliferation of digital multimedia data creates a technological revolution to the entertainment and media industries and introduces the new concept of web-based social networking communities. However, the massive production and use of digital media also pose new challenges to the scalable and reliable sharing of multimedia over large and heterogeneous networks, demand effective management of enormous amount of unstructured media objects that users create, share, distribute, link and reuse, and raise critical issues of protecting intellectual property of digital media data. The proper management and protection of digital multimedia at such an unprecedented scale are beyond the capability of current technologies and demand new solutions. This collaborative research effort between University of British Columbia and University of Alberta tackles the emergent technical challenges (e.g. information management and content protection) in large-scale media social networks. The aim is to establish a multimedia management and security framework to provide effective management, secure and reliable sharing of digital media in large-scale social networks. In particular, this talk addresses a summary of our recent research efforts in the following areas: content-based fingerprinting for media indexing and content recognition; understanding and analyzing the impact of human factors on multimedia systems; and building an automated network-service monitoring paradigm.

Dr. Darko Kirovski of Microsoft Research, Redmond gave the next talk of "Realizing the Uniqueness of Optical Media." When a DVD is stamped it is physically unique. He proposes a scheme to detect and deploy this uniqueness for the benefit of Digital Rights Management.

The next talk of "Connectivity and Security in Directional Multimedia Sensor Networks" by Prof. Deepa Kundur of Texas A&M University discussed the recent increased interest in the development of untethered sensor nodes that communicate directionally via directional radio frequency (RF) or free space optical (FSO) communications. Directional wireless sensor networks, such as the original Smart Dust proposal that employs broad-beamed FSO communications have the potential to provide gigabits per second speeds for relatively low power consumption suitable for multimedia sensing systems. Two significant challenges shared by the class of directional networks are connectivity and routing security, especially for random deployments. In

this talk, two issues are addressed: 1) the feasibility of employing directional communications paradigms in large-scale security-aware broadband randomly and rapidly deployed static multimedia sensor networks; 2) the implications of link directionality to network connectivity and secure ad hoc multihop routing and highlight approaches in network design to mitigate compromising between the two.

The talk of "Bounds on Biometric Security" by Dr. Ton Kalker of Hewlett Parkard gives an overview of some recent work on trade-offs between the capacity and security of biometric systems. He shows that it is possible to formulate bounds for a number of cases, and that some of the classical schemes (for example fuzzy commitment) are sub-optimal. He also sketches many open questions. A highly entertaining presentation, entitled "Cognitive Sensors Networks: The New Frontier for DSP," was given by Prof. Magdy Bayoumi of University of Louisiana at Lafayette. Computers, communication, and sensing technologies are converging to change the way we live, interact, and conduct business. Wireless sensor networks reflect such convergence. These networks are based on collaborative efforts of a large number of sensor nodes. They should be low-cost, low-power, and multifunction. These nodes have the capabilities of sensing, data processing, and communicating. Sensor networks have a wide range of applications, from monitoring industrial facilities to control and management of energy applications, to military and security fields. Because of the special features of these networks, new network technologies are needed for cost effective,low power, and reliable communication. These network protocols and architectures should take into consideration the special features of sensor networks such as: the large number of nodes, their failure rate, limited power, high density, etc. Moreover, applications and impact of Sensors Networks are going to a higher and wider levels through the development of cognitive capabilities of these networks. Cognitive Sensors Networks, CSN, represent a transformational impact on technologies, applications, and expectations. In this talk the impact of wireless sensor networks will be addressed, several of the design and communication issues will be discussed, and a case study of a current project of using such networks in drilling and management off-shore oil and natural gas in the gulf region are given. The main criteria, expectations, and objectives of CSN are also highlighted

One special presentation was made on the topic of "The H- INDEX," by Prof. Yucel Altunbasak of Georgia Tech, with open disucssions. A while ago, with the help of some Ph.D. students, the presenter compiled H-indices for image processing researchers. He received several feedback about publishing this list. The most common feedback were: 1) How do you define an image processor? 2) How do you resolve the different H-index characteristics of image processing, computer Vision, and computer-science oriented people? 3) ISI-based and Google-based H-index numbers are vastly different. Which one do you use?, and most important of all, 4) what is the use of publishing such a list? Would it benefit the society? Further feedback from the society is still needed regarding 1) if publishing such a list (online) would benefit the society or be harmful because of possible misinterpretation (esp. by students), and 2) how best to do this project.

3 Outcome of the Meeting

The workshop II follows the workshop I over 4 years ago, and provides a timely and updated cross-disciplinary bridge among the relatively new area of multimedia, the well-established discipline of mathematics, and the emerging area of machine learning that heavily depends on mathematics. For many researchers in a specific area of multimedia, the workshop provided an excellent opportunity to broaden their perspective, exceeding the level achieved 4 years ago. A series of the workshop's high-quality presentations made clear the surprisingly similar mathematical and machine-learning approaches applied to speech, audio, image, and video-processing research. The presentations and intensive discussions enabled participants to examine the variety of approaches in different media areas, an invaluable opportunity made possible by the mixed forma and linformal styles of the workshop.

We would like to have BIRS to continue sponsoring cross-disciplinary workshops such as the series of two that we organized. Cross-disciplinary research sharing similar mathematical approaches, which now expand significantly to machine learning approachesm, stands to benefit the most from such workshops. The different branches of media processing research make it impossible to gain expertise in every sub-area, and our BIRS workshops helped immeasurably to foster an awareness of new trends in the various sub-disciplines. This is particularly important to some industrial researchers whose work has a relatively short-term scope. This was the case 4 years ago, and not is still the case although to somewhat less extent.

Most researchers in multimedia cannot afford the time-consuming process of mastering the subtleties of all the multimedia processing techniques. Our BIRS workshop provided an ideal opportunity to make close connections among them and to deepen our understanding of problem areas. The workshop succeeded in its aim to bring mathematicians, machine learning researchers. engineers, and scientists to interact and get exposed to each others ideas and advances in these disciplines. As different multimedia technologies have evolved and continue to evolve at a very rapid rate, the exact definition of multimedia remains illusive, even though multimedia technologies are now being widely deployed in industries in a multitude of applications.

The cross-fertilization among the different disciplines, academics and practitioners, engineers and mathematicians encouraged by the workshop was very useful in exposing the different communities to a new range of challenging and timely technical advances, the underlying mathematical problems and applications, and implementation challenges. This workshop II advances what was achieved by Workshop I by expanding the mathematical approaches to include the most relevant machine learning aspects. This expansion is particularly beneficial for multimedia research because of its cross-disciplinary nature and because of the intricacy on many of its sub-areas.