Surviving the data deluge: Scalable feature extraction, discrimination and analysis

Svetha Venkatesh, Dinh Phung and Duc Son Pham

Background and experience of participant in field

Venkatesh holds the Inaugural John Curtin Distinguished Professorship and the Founding Director of the Research Institute for multisensory Processing and Content analysis, IMPCA, with a proven ability to inspirationally build and lead successful multidisciplinary teams and strong international collaborations. Her track record indicates her ability to address substantial pure and applied research problems focusing on large scale pattern recognition. This includes new techniques for scalable multimedia analysis, new theoretical formulations for scalable probabilistic models and new paradigm shifting methods for wide area surveillance. Venkatesh has developed frontier technologies in large scale pattern recognition exemplified through 387 publications including 2 books, 12 book chapters, 87 journal and 276 conference publications. She has been involved in two start-up companies: Virtual Observer is based on the leverages mobile cameras to deliver wide area surveillance solutions and won the Runner up in both the *WA Inventor of the year (Early stage) and Global Security Challenge* (Asia-Pacific) in 2007. A second start-up company *iCetana* started in 2009, that deploys novel techniques for large scale video anomaly detection and won the Broadband Innovation award at TECH23 in 2010. Venkatesh was elected a *Fellow of the International Association of Pattern Recognition* in 2004 for contributions to formulation and extraction of semantics in multimedia data. She is a *Fellow of the Australian Academy of Technological Sciences and Engineering.*

The vision of the participant:

The world is awash with data from proliferating distributed and multimedia sensors. An International Data Corporation whitepaper[1] (March 2008) notes "the digital universe in 2007 was 281 billion Gigabytes" and for the first time exceeded available storage. It predicts that by 2011, more than half of this information will not have a permanent home. Surveillance cameras, sensor-based applications, and social networks are among the named drivers of this explosion. We propose avenues of research to address the underlying issues in the collection and analysis of data from pervasive, heterogeneous and distributed sensors.

Sensing: With the increasing sensors, comes the problem of information extraction. The machinery to jointly analyse this collection of sensors lacks scalable approaches. One place to fundamentally re-examine the *scalability* issue is the process of *sensing* itself. Recent research in compressed sensing has exciting results on how sensing itself can be done in a *compressed* way (not to be confused with compression of stream itself). This means that if there were million sensors which were being jointly analysed, compressed sensing would provide opportunity to sample *a selection of the sensors*. Significantly, the "compressed signal" sensed can recover the original signal (in this case, as if all the sensors were sampled) with high probability. This opens many fields of exploration: How can we apply compressed sensing across millions of sensors to capture data directly in an information theoretic way? What are the ways in which we can exploit compressing sensing with pervasive devices? How can we directly analyze the compressed sensed data?

Analysis: Much past work on analysis had focused on analysis methods that build models using *supervised* data or *supervised model parameter setting*. These two core issues prevent scalability to adapt to new and large data streams, which inherently require models to adapt dynamically to changing data streams. Recent Bayesian nonparametric methods in machine learning go toward addressing this challenge. Importantly the model complexity adapts as the data comes in. This opens up many fields of exploration: Can we explore unsupervised

[1] www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

methods for controlling parameter growth, modeling dynamically changing data streams from large numbers of sensors? How we can exploit the inherent property of these algorithms to adapt to the scale and robustness required for pervasive sensing?

Multi-scale decision making: Decision making with data collected from disparate and many sources is difficult. Traditional methods have relied on methods to evaluate the probability of events and models to combine these partial results. Recent work in psychology explores the strong role of affect in decision making, in particular its role in being able to navigate situations with uncertainties. There are two kinds of affect: expected emotions (evaluated by the person on the basis of outcome chosen) and immediate emotions (current user state). The role of affect in decision making in an algorithmic sense is largely unexplored. We see a clear link in pervasive computing and immediate emotion detection and the larger role of both types of affect in decision making.

We propose the choice of a domain to study this problem, and we choose the field of pervasive assistive techniques for autism as a starting point. Affect and its regulation are difficult for people with autism. We propose the building of new sensors and technologies for pervasive detection of mood and its incorporation into devices for decision making. Questions include: What kinds of new sensors do we need? How do we build sensory devices through which an autistic person can learn to regulate affect? How can we use these insights to build useful decision making algorithms in other domains?

Evidence that pursuing this vision will lead to major advances:

We argue that each of the sub-areas we have chosen have exciting new developments in theory in an allied discipline. There is therefore strong theory as the foundation, upon which new algorithms and applications can be built. Importantly, each new development opens new ways of transcending barriers current algorithms face for scalably adapting and coordinating among a large number of sensor streams. We outline the specific recent development in each sub-area.

Sensing: The current sensing paradigm, a legacy of Nyquist, requires sampling at twice the highest frequency, dooming us to compress after capture. Much of what we sense is immediately thrown away. Breakthroughs in signal processing called *Compressed Sensing (CS)[1]* shows that it is possible to preserve the information with few non-adaptive linear measurements whose sampling rate is only proportional to the *information level*. This opens up a plethora of possibilities for information extraction across millions of pervasive sensors.

Analysis: The growing interest in non-parametric techniques in machine learning [2] has opened a rich and scalable way to explore large data streams. The formulism provides a paradigm to grow the model complexity as more data comes in, not requiring an explicit training phase, or fixed parameter choices. The last two aspects make it a foundational technology to consider large scale data streams, and much work remains to be done.

Multi-scale decision making: Recent work in psychology explores the strong role of affect in decision making [3]. There is immense scope to explore how algorithms can be constructed to learn about and from affect, and how it can be used to modulate multisensory fusion. We chose pervasive assistive devices for autism. 1 in 150 children has Autistic Spectrum Disorder (ASD showing impairment in social interaction, communication, cognitive functioning, and adaptive behaviours. The US National Standards Report in 2009 [NAC 2009] estimated a societal cost of $3.2 million during the lifespan of an ASD individual. The impact for pervasive assistive devices in thus immense.

1. D. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
2. Y.W. Teh and M.I. Jordan, Hierarchical Bayesian Nonparametric Models with Applications**,** *Bayesian Nonparametrics*, Cambridge University Press, 2010.
3. Loewenstein, G., & Lerner, J.S. (2003). The role of affect in decision making. In Davidson R.J. et al. (Ed.), *Handbook of Affective Sciences* (pp. 619--642). Oxford New York: Oxford University Press.