Frank Nack
University of Amsterdam, nack@uva.nl

Content is Dead . . . Long Live Content: The New Age of Multimedia-Hard Problems

Lexing Xie
Australian National University

David A. Shamma
Yahoo! Research

Cees Snoek
University of Amsterdam

In multimedia research, content analysis has always held a major role—burgeoned on a foundation of machine learning and data-intensive algorithms. Yet the past decade has given rise to rich user annotations and massive multimedia sharing. Social multimedia computing quickly moved beyond computing features to ultimately change how we understand media semantics by adding into the equation user-generated tags as well as social networking, location-based services, and mobility (or SoLoMo, as it is often abbreviated). Classic multimedia problems seem to be simplified because it is so much easier to automatically identify, for example, a picture of the Golden Gate Bridge if you have a mostly accurate location information in the metadata.

So is there any role for the still iterating success of visual analysis in finding that bridge photo? What is the future of content analysis in the world of annotation, tags, and crowd wisdom? We hosted a panel at ACM Multimedia 2012 in Nara, Japan (see Figure 1). Four panelists (Susanne Boll from the University of Oldenburg, Tat-Seng Chua from National University of Singapore, Minoru Etoh from NTT DoCoMo, and Malcom Slaney from Microsoft Research) and an audience from academia and industry debated whether content can continue to play a dominant role in multimedia research, if it has become secondary, or if it is dead in the age of social, local, and mobile media.

Editor's Note

Following the discussion of the ACM Multimedia 2012 panel on content analysis, the authors further investigate whether content can continue to play a dominant role in multimedia research in the age of social, local, and mobile media. In this article, they propose that the community now must face the challenge of characterizing the level of difficulty of multimedia problems to establish a better understanding of where content analysis needs further improvement. They also suggest a classification method for multimedia problems.

"Multimedia-Hard" Problems

Over the course of the ACM Multimedia session, the panelists kept revisiting what they referred to as hard and easy multimedia problems, which are referred to as two distinct problem classes.

The panel remarked on several well-known successes of artificial intelligence algorithms such as face detection, optical character recognition, music fingerprinting, and speech recognition. Each of these widely used algorithms now enable a range of applications that uses them as components. In other words, multimedia applications using these components can now be seen as easy problems, such as voice-driven Web search.

The panel then discussed a perspective on the scope of multimedia problems and remarked that multimedia isn't necessarily about media, but about problems that can be solved with input from multiple sources. Leveraging context to answer questions about media content was given as one concrete example. With media content available in social networks, location-based services, and mobile devices (SoLoMo), the problem of content understanding transforms into the joint understanding of the location or personal context and content. For example, although general recognition of an urban scene is a difficult problem, recognizing buildings and landmarks given a GPS location becomes solvable because it drastically narrows the search space. Another example is to let human computations solve the problem. For instance, Duolingo uses the output from online language-learning

1070-986X/14/$31.00 © 2014 IEEE

Published by the IEEE Computer Society

Classifications of Computational Problems

Our notion of multimedia-hard (MM-hard) builds off of key terminologies from problem classification in computational complexity[1] and AI. NP (nondeterministic polynomial time) is the set of decision problems where the "yes" instances can be accepted in polynomial time by a nondeterministic Turing machine. Informally, a problem is NP-complete if it is NP and as "hard" as any problem in NP. One common practice of showing a problem is NP-complete is first to show that it is NP and then to reduce some known NP-complete problem to it. NP-hard problems are, informally, at least as hard as the hardest NP problems. A problem *L* is NP-hard if it is at least as hard as an NP-complete problem, but *L* does not have to be in NP.

This problem classification terminology found its way into artificial intelligence in the early 1990s. The terms AI-hard or AI-complete[2] are used to describe the difficult problems in AI. The term "complete" here is used to describe a nontrivial replication of human intelligence, aligned with the strong AI paradigm.[3]

References

1. T.H. Cormen et al.,*Introduction to Algorithms,* 3rd ed., MIT Press, 2009.
2. E.D. Raymond, *Jargon File,* version 2.8.1, 22 Mar. 1991.
3. J. Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences,* vol. 3, no. 3, 1980, pp. 417–457.

Figure 1. Panelists and audience during the panel debate at ACM Multimedia 2012.

activitiesto translate Wikipedia and other documents to many languages (see http://en.wikipedia.org/wiki/Duolingo). The infusion of human intelligence into the content analysis process is achieved by either a community of people or a group of crowdsourcing workers. Such human-in-the-loop processes will help in both understanding content and disambiguating context.

These two examples speak to the roles people play in multimedia in today's world. More formally, it introduces people into the computational flow. To date, we have not attempted to characterize the difficulty of multimedia problems. The goal of this article is to go beyond the debate that occurred at ACM Multimedia 2012 and to propose such a classification. (See the "Classifications of Computational Problems" sidebar for related classification research.)

Defining MM-Hard Problems

We attempt to describe multimedia problem difficulty in a way that resembles problem classification in AI and is inspired by human-assisted computation themes. Similar to AI, we use the term multimedia-hard (MM-hard) to describe

MM-hard refers to multimedia problems that require human-level insights and perception that can't be realized with a single algorithmic approach.

difficult problems in multimedia. More formally, MM-hard refers to problems in multimedia that require human-level insights and perception that are yet incapable of being realized through a single algorithmic approach. Thus, it may require a humans-in-the-loop aproach, where the system makes use of crowdsourcing workers or uses user-generated production, annotation, or dissemination of media. With obvious term similarities to the strong AI hypothesis,[1] we assert that the distinction in multimedia research lies within clear semantic representations that consider the implications of perception and exhibit multimodal reasoning.

This notion of "strong" is attractive because it appeals to the primary goal of machine intelligence research. A more recent formalization of AI problems is the notion of human-assisted Turing machines (HTM).[2] An HTM is a Turing machine with access to an oracle, or human *H*. Here a tuple $<\Phi_H(M), \Phi_M(M)>$ is used to denote the complexity of HTM *M,* where $\Phi_H(M)$ is the time complexity required from the human oracle and $\Phi_M(M)$ is the time complexity of the conventional Turing machine. The HTM is a promising formalism for multimedia problem classification for three reasons:

- There is an AI component in many multimedia problems, such as image/video/audio recognition, tagging, and search.
- The HTM framework can account for uncertainty in the human oracle or can be extended to account for multiple human oracles.
- Having complexity classes for multimedia problems will enable problem reduction—that is, more easily transfer solutions from one problem to another—or define new problem classes.

In essence, solutions to MM-hard problems require such an oracle, be it an individual, community, or crowd, for computation. Beyond the limitations of time and space, the problems multimedia faces moving forward are not a function of pixels or audio samples; they require insight into how media is captured, shared, and manipulated. This begins with expanding our existing frameworks and formalisms.

Problem Difficulty and Reduction

The notion of MM-hard has the potential to benefit multimedia research in two fundamental ways. The first is to describe problems in terms of their (machine and human) difficulty; the second is to be able to do problem reduction—that is, convert one problem to another and compare problems.

One of the building blocks to achieving this is a partial ordering for the difficulty of problems. For example, text to speech is considered a solved problem with efficient and high-quality polynomial-time algorithms,[3] optical character recognition can be considered an HTM $<O(1)$, polyonmial$(n)>$ for recognizing n symbols from a fixed-size vocabulary,[2] and classification of n samples is $<O(n)$, $O(n)>$ if the human oracle sees all input and classifies it or $<O(\log n)$, $O(n \log n)>$ if the machine sorts the samples and queries the human oracle for a suitable threshold.[2]

The other building block consists of reduction methods to convert one problem to another. For instance, speech understanding can be reduced to a question-answering problem using text to speech as one of the conversion components.[4] For the multimedia domain, a visual matching problem with a fixed (but large) vocabulary, such as trademark recognition, can be reduced to an optical character recognition problem that requires $O(1)$ human time.

The HTM construct naturally permits extensions for the complexity classes to account for a diverse set of problems that arise in real-world scenarios. Examples can include nondeterministic HTMs (for example, analogous to nondeterministic TMs) that consist of decision problems that can be verified by a deterministic HTM in poly-time, such as common-sense planning tasks.[2] An HTM can also permit probabilistic output from the human oracle or parallel and distributed computation in both machines and humans. Such parallel and distributed variants have had important

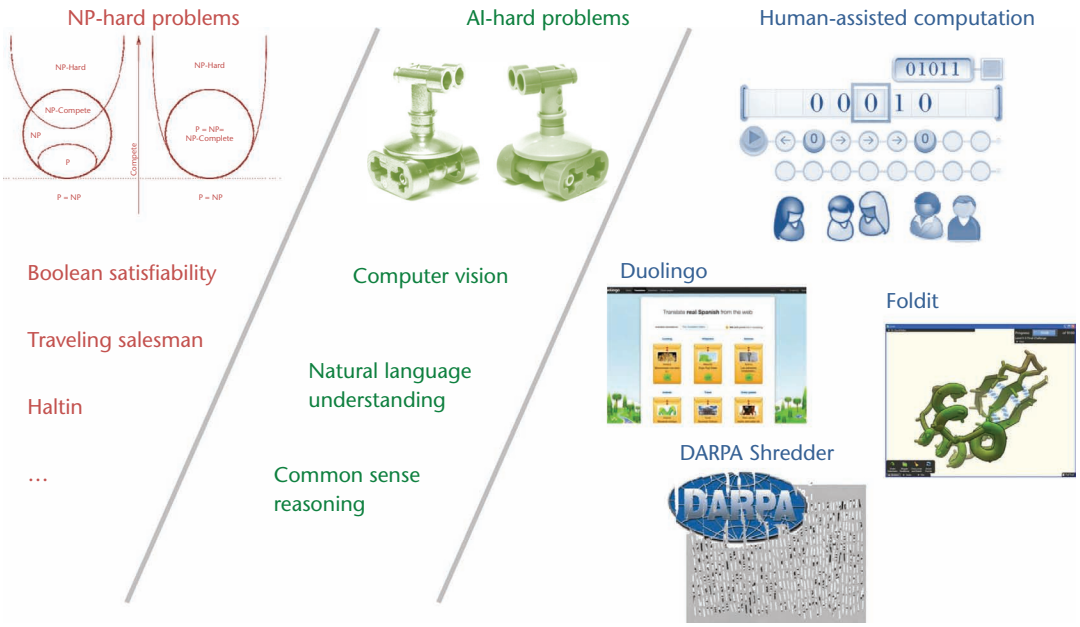

Figure 2. An overview of three problem classification schemes and their example problems. (a) P versus NP and example NP-hard problems. (b) Example AI-hard problems. (AI Photo by Don Solo, CC BY-NC-SA 2.0, http://www.flickr.com/photos/donsolo/3302526343.) (c) Human-assisted computation examples include the Duolingo online language learning and crowd-sourced translation application; the FoldIt crowd-sourced protein folding game;[5] and the 2011 DARPA Shredder Challenge (http://en.wikipedia.org/wiki/DARPA_Shredder_Challenge_2011), with the winning entry employing a human-in-the-loop approach with computers suggesting likely solutions.

applications and significant impact in recent years, such as distributed content creation (including the online encyclopedia Wikipedia and question-answering sites Stackoverflow and Quora) and crowd-sourced energy minimization problems in biochemistry such as FoldIt[5] (see Figure 2).

An Outlook for MM-Hard and Human-Assisted Computation

One may ask, what does MM-hard mean? Does MM-hard change the "semantic gap"?[6] The answer to this important question is twofold. First, MM-hard allows us to quantify the semantic gap. Explicitly knowing that some problems are harder than others can help direct problem-solving efforts. Second, an MM-hard approach lets us transfer and extend known solutions. Being able to say, for example, that video copy detection, music identification (such as the Shazam Music Identification Service), and optical character recognition are problems of roughly the same difficulty (subject to computational resources) can be a first step toward adapting the solution from one problem to another. Although MM-hard problem classes do not close the "semantic gap" themselves, it is useful to know which problem are facing a bigger semantic gap, when applicable.

It is also worth noting that the landscape of known problems are evolving and shifting, thus changing the meaning of a "semantic gap." Speech recognition, for example, has been a long-standing research challenge for several decades, and the term "ASR complete" was coined to describe its variants.[7] It was only recently that various systems has achieved high enough recognition accuracy to allow real-time human-computer interaction for conversational speech in natural environments (such as Apple's Siri).[8] Another such example is a subset of question-answering problems that is free of memory and context; this too is becoming one of the solved problems with the success of IBM's Watson.[9]

Finally, there are new problems and applications emerging in the spaces between problems of known complexity classes, especially considering the different behaviors of the many human oracles acting in a distributed and probabilistic fashion. One such example can be a "greedy" human oracle in interactive

applications. Take photo album organization, for instance; a user knows whether or not one layout is better than the other but does not know the governing criteria that will generate a global optimal. Another example of a challenging problem can be information routing on real-time social networks, for example, to relay a message via Twitter for a small-world experiment.[10] This amounts to a networked, online, streaming version of a decision problem—every person on the route will need to decide who to route the message to, yet the message will also compete with other messages from the network neighborhood.

There are a number of open challenges before MM-hard can be claimed as the framework for generic problem description and reduction. Just as there is no known catalog of all AI problems in terms of their complexity, building such a MM-hard catalog may require many limiting assumptions. It will still be valuable, however, to catalog as many challenges as the problem nature allows, because doing so will have immediate benefits for complexity comparison and reduction.

Concluding Thoughts

Content understanding problems have been and will continue to be a major part of multimedia research. Content, context, and human-powered computation have emerged as research themes in multimedia analysis, as reflected in the panel discussions at ACM Multimedia 2012. We have used this article to relate media analysis problems to known complexity classes of computation, AI, and human-assisted computation. We believe that such a problem structure will benefit algorithm research and help researchers develop novel applications that can be reduced to subsets of known problems. **MM**

Acknowledgments

We thank the MM 2012 panelists (Susanne Boll, Tat-Seng Chua, Minoru Etoh, and Malcolm Slaney) for their vision statements and inspirations that led to this article. Special thanks goes to Susanne Boll for sharing her recollections about the discussions, and Frank Nack for his constructive editorial suggestions. We also thank Yong Rui, Shih-Fu Chang, Patrick Pérez, and Wei Tsang Ooi for discussions. Last but not least, our appreciations go to the enthusiastic panel participants for a lively debate.

References

1. J. Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences,* vol. 3, no. 3, 1980, pp. 417–457.
2. D. Shahaf and E. Amir, "Towards a Theory of AI Completeness," *Proc. AAAI Spring Symp.: Logical Formalizations of Commonsense Reasoning,* AAAI, 2007, pp. 150–155.
3. W.B. Kleijn and K.K. Paliwal, *Speech Coding and Synthesis,* Elsevier Science, 1995.
4. R.V. Yampolskiy, "AI-Complete, AI-Hard, or AI-Easy: Classification of Problems in Artificial Intelligence," tech. report, Dept. of Computer Eng. and Computer Science, Univ. of Louisville, Sept. 2011.
5. S. Cooper et al., "Predicting Protein Structures with a Multiplayer Online Game," *Nature,* vol. 466, 2010, pp. 756–760.
6. A.W.M. Smeulders et al., "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, no. 12, 2000, pp. 1349–1380.
7. N. Morgan et al., "Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), vol. 4, IEEE, 2003, pp. 740–743.
8. G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: the Shared Views of Four Research Groups," *IEEE Signal Processing,* vol. 29, no. 6, 2012, pp. 82–97.
9. D. Ferrucci et al., "Building Watson: An overview of the DeepQA Project," *AI Magazine,* vol. 31.3, 2010, pp. 59–79.
10. J. Travers and S. Milgram, "An Experimental Study of the Small World Problem," *Sociometry,* vol. 32, no. 4, 1969, pp. 425–443.

Lexing Xie is a senior lecturer and fellow of computer science at the Australian National University. Contact her at lexing.xie@anu.edu.au.

David A. Shamma is a senior research scientist and head of the HCI Research group at Yahoo! Research. Contact him at aymans@acm.org.

Cees Snoek is an associate professor at the University of Amsterdam and head of R&D at Euvision Technologies. Contact him at cgmsnoek@uva.nl.