Overcoming Barriers to Progress in Multimodal Fusion Research

John S. Garofolo

National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899
john.garofolo@nist.gov

Abstract

Government programs have often led the way in research in information detection, recognition, and extraction technologies. Significant performance improvements have been achieved over the last 20 years in automatic speech recognition, spoken language understanding, text, image, and video retrieval, textual structural and semantic extraction, machine translation, speaker recognition, language recognition, face recognition, video person/object detection and tracking, and other extractive technologies. Performance peaks are being observed in these technologies and the pace of breakthrough algorithmic improvements is slowing. Initial efforts at putting these technologies together into complex applications have involved pipelining with progressive distillation or time lining with limited success. It has become clear that the tightly-coupled fusion of these technologies is the next logical step in the progression towards human-like computational intelligence capabilities. Unfortunately, there have been both multi-disciplinary and technical barriers to research and development in the area of multimodal fusion and it has yet to gain significant momentum. The gargantuan amounts of multimedia now appearing on the Internet and elsewhere are necessitating an acceleration of work in this area. We identify challenges that must be overcome to enable critical development and speed progress in multimodal fusion technologies.

Multi-Disciplinary Challenges

R&D Focus – The ubiquity of information technology, cell phones, digital cameras, and the Internet have evolved very quickly over the last 10 years. It would have been difficult in 1998 to envision the amount of multimedia data that currently resides on the Internet. Organizations charged with processing these data are still trying to catch up to the information revolution that has taken place. Many unimodal extraction technologies remain immature and continued research is necessary to perfect them. As such, they have remained a focus for R&D efforts. Unfortunately, however, with a few notable exceptions, there have been few efforts addressing research in multimodal fusion technologies – certainly far less than for individual human language technologies. As a result of this imbalance, integration has remained largely a post-research process -- often relying on loosely-coupled pipelined components. Complex multi-component systems developed in this way tend to suffer from compounded errors. Relatively small errors in components can cause error cascades as they are multiplied by successive imperfect processes. While individual processes may have what seem to be relatively low error rates, overall system performance can be significantly degraded.

Increased research in multimodal fusion will expedite the creation of synergistic solutions to important multimedia extraction challenges that will reduce the error cascading effect as well as produce more cohesive information across modalities. Complex integration efforts must address this challenge through information-based solutions that span modalities, domains, and component technologies. Multi-disciplinary as well as technological boundaries must be crossed to address the challenges since researchers must begin to design their algorithms to fit into the "big picture". As such, there must be a significant amount of communication in the R&D process across areas of expertise.

Critical Mass – While a number of workshops and conferences that address multimodality have been active for some time, the number of researchers focusing on multimodal fusion and the number of new researchers in this area has not grown commensurately with the vast quantities of multimedia data that are being generated.

Because of the complexity in creating fusion applications and because these challenges call for the development of "renaissance" researchers, a significant community-building effort needs to be undertaken to create and increase expertise in this area. Cross-domain knowledge must be grown in speech, vision, and natural language processing technologies, and in behavioral and social science. To accelerate the development of the research community, new educational resources need to be created to encourage students and young researchers to gravitate toward this area. Such resources include cross-domain seminars, shared data, shared algorithmic components, and fusion frameworks. These resources combined with a focus on fusion research and applications will help to build a researcher base whose specific expertise is in multimedia information extraction.

Technological Challenges

Data Resources and Evaluation – While multimedia data have been mined to create research corpora for unimodal applications research, very little multimodal research corpora currently exist. Evaluation methodologies have previously focused on either component technologies or system usability. Significant infrastructure challenges will need to be overcome to enable the progressive development of multimodal fusion technologies.

Significant new resources must be created to enable research, development, and evaluation in multimodal fusion technologies. These include the development of large diverse datasets with complex annotations. The good news is that a great deal of multimedia source data already exists and is readily accessible. However, the reference annotations necessary to make these data useful for multimodal research are lacking. Such annotation data will be challenging and expensive to create since they will necessarily span modalities. Synchronization will be critical to maximizing the utility of these data. Metadata standards and practices are a critical component that must be addressed since the structures for representing and sharing multimodal information will be complex. Commonalities must be identified and standards must be created to enable reuse and to make the development of these resources cost effective. Such standards must necessarily cross technological boundaries.

A significant body of work has been created over the last 25 years in the evaluation of component technologies and in usability analysis of end-to-end systems. The evaluation of hybrid technologies presents new challenges. While most simplistically these technologies may be treated as a set of extraction tasks, effective evaluation methodologies will provide important failure analysis feedback. Such feedback will be difficult to extract from tightly coupled component technologies. New metrology research is required to address this challenge.

Evaluation effectiveness will be highly dependent on the ground truth that is used to assess algorithm performance. These data must be more accurate than the automatic algorithms that are being tested for evaluation to provide statistically significant results. Given the complexity of multimodal/multimedia tasks, such ground truth will be difficult to create with the necessarily high level of accuracy. New more sophisticated annotation tools that perform much like multimedia analyst tools will need to be created to assist human annotators in creating gold standard information for evaluation. New annotation methodologies will likewise need to be created that help to minimize human error/uncertainty in the evaluation results.

Core Fusion Research – Pipelining and progressive distillation are unfortunately often misinterpreted as fusion. These currently prevalent approaches suffer from cascading errors, missing information, and unresolved discrepancies. Fusion requires early coupling of extraction components that work synergistically towards the creation of information. Multimodal fusion techniques will render these loosely-coupled approaches obsolete for multimedia information extraction.

In order for multimodal fusion techniques to be realized, research breakthroughs need to take place so that parallel heterogeneous data sources can be processed synergistically. New research must be carried out to enable the creation of the necessary data-sharing structures between tightly-coupled unimodal components. Extensive research must also be performed in choosing what to fuse and how to fuse it and how to most create the most effective multimodal hybrids. Processing speed and computational complexity will be an issue as component technologies are fused. A combination of novel statistical methods, knowledge and data representation methods, heuristics, and computationally efficient algorithms will be required to create efficient reusable fusion solutions.

Driver Applications – Multimodal fusion research has previously lacked pull from high-impact application needs to motivate and guide research directions. Until recently, this area has suffered from a "chicken and egg" problem as technology customers have struggled with envisioning new ways of working with multimedia data and researchers have struggled with divining the types of multimodal tools that would be useful for these customers.

Given the current prominence of multimedia data on the Internet and the creation of early tools to manipulate this data, we can now begin to envision a variety of useful applications. Customers are beginning to talk to researchers about application needs. Key driver applications now need to be defined with both customer and researcher input that will help propel the research forward and create revolutionary multimodal applications.

Summary

Both multi-disciplinary and technical barriers need to be overcome for multimodal fusion technologies to be effectively brought to bear on the ever-increasing deluge of multimedia data being created daily on the Internet and in other important domains. Critical mass will be created through an increased focus on multimodal fusion research and research community development. Reusable multimodal data resources and cross-cutting standards need to be created to make development of these technologies cost effective. Novel evaluation methodologies need to be created that are supported by revolutionary methods in gold standard annotation creation to support progressive development. Key research needs to take place both in the creation of fusion constructs and development of effective hybrid technologies. This research will be most effectively motivated through key high-impact driver applications that are created as the product of a dialogue between the research community and technology customers.