Seeing is Believing:
The Quest for Multimodal Knowledge

Gerard de Melo[1] and Niket Tandon [2]
[1]IIIS, Tsinghua University
[2]Max Planck Institute for Informatics

There is a growing conviction that the future of computing will crucially depend on our ability to better exploit data to produce more intelligent systems. Increasingly, this will involve drawing simultaneously on multiple heterogeneous modalities, to take full advantage of the vast quantities of images and videos now available on the Web and elsewhere. We give several examples of methods that leverage prior knowledge for better, more semantically informed visual analytics, as well as methods that use multimodal data for better textual analytics. Important progress may come from approaches specifically geared towards harvesting rich multimodal knowledge. For example, our Knowlywood system relies on Hollywood movies to learn about human activities. Once acquired, knowledge of this sort can then be re-used across different tasks, much like humans draw on their accumulated knowledge when making sense of the world.

DOI: 10.1145/2903513.2903517 http://doi.acm.org/10.1145/2903513.2903517

1. INTRODUCTION

In the past few years, we have seen remarkable new achievements in developing intelligent systems. For instance, humans can no longer keep up with IBM's question answering system Watson, not even all-time Jeopardy! champions such as Brad Rutter and Ken Jennings. In speech recognition, markedly lower error rates have enabled powerful dialog systems, including Siri on Apple's iOS, Alexa for Amazon Echo, and various others powering advanced customer support services. In image object detection, the error rates in the prominent ILSVRC competition have dropped by orders of magnitude in the past years, reaching human-level results in certain task setups.

Two major factors that have driven much of this progress are 1) the availability of massive amounts of data and 2) algorithmic advances such as deep learning. Both of these will continue to propel us ahead in the coming years. Yet, the path ahead is not straightforward.

From a data perspective, although we can now collect massive amounts of data on the Web and elsewhere, this data comes in a multitude of different forms that cannot easily be plugged into existing algorithms or systems. Regular supervised machine learning, for instance, requires labeled training data that tends to be very costly to procure. In natural language processing, for instance, the Penn Treebank training corpus was produced by human annotators over the course of eight years from 1989 to 1996 [Taylor et al. 2003].

While sufficient for traditional statistical NLP methods of the past, it turns out that deep recurrent neural networks need even more training data [Vinyals et al. 2015]. In the coming years, it appears that we will need to find novel ways to take advantage of the different kinds of data that are available.

From an algorithmic perspective, current machine learning algorithms excel at mathematical optimization and at discerning discriminative patterns in large amounts of labeled training data. While deep learning has enabled us to learn more sophisticated models than in previous work, most of them are particularly data-hungry, requiring massive amounts of labeled examples to learn a new category. This is still quite different from the abstract mental models that humans are able to form given just a single example of a new concept.

In the future, it appears, we will need to devise more advanced models capable of genuinely modeling the world. A system of this sort should be able to learn to recognize a mime performing on the street given just a single training image and description. Identifying discriminative visual patterns from just one example is a challenge for current algorithms, but may be possible if the system can benefit from rich prior knowledge as background information. Equipped with such a richly informed model, a system may be able to infer that mimes are humans, that what appears to be their face seems to have strikingly white makeup (different from average humans), and perhaps that what appears to be their clothing seems to have striped patterns.

Hence, overall, obtaining a deeper understanding of the world will require new models that simultaneously draw on multiple modalities and multiple heterogeneous forms of Big Data to better organize information and produce more intelligent systems. While we are still far from such advanced models, this paper presents possible avenues towards systems that learn from multiple modalities and collect multimodal knowledge. The hope is that this will lead to collections of general-purpose knowledge that can then be reused across different tasks, much like how humans draw on their rich prior experience in the world when faced with a new concept or task.

2. VISUAL ANALYTICS BY LEARNING FROM HETEROGENEOUS DATA

Computer vision has traditionally often been viewed as a mere pattern recognition problem. Deep convolutional neural networks have enabled us to learn higher-level image representations and distinguish among the 1000 ImageNet object categories with very high accuracy. Still, there is some way to go until we have systems that can analyse a scene and fully interpret it, with a detailed segmentation and fine-grained labels. An ideal system would be able identify that a historic photo portrays John F. Kennedy with his bride Jacqueline sitting at a table and eating pineapple salad. To move towards such capabilities, we will need to draw on both visual cues and on prior knowledge to develop more elaborate models of what is being perceived. This entails exploiting multiple heterogeneous forms of data.

2.1 ShapeLearner/ShapeExplorer

The ShapeLearner/ShapeExplorer system [Ge et al. 2016] is one example of a system that attempts to draw on background knowledge for more fine-grained image understanding. The system jointly recognizes objects together with a hierarchy of their parts. For instance, it not only identifies a horse but also its legs and head, and their corresponding parts, e.g. the nose and mouth on the head.

While most computer vision systems focus on classifying only entire images or rectangular bounding boxes, ShapeExplorer yields a detailed 2D segmentation of the image. Rather than generic segment labels such as foreground/background, or sky/water/etc., each segment is labeled with a fine-grained object or part label.

When learning new categories, the system draws on its familiarity with related categories, as revealed by a taxonomic hierarchy of categories. For example, when learning about zebras, much of the prior knowledge that the system has about horses can be reused.

2.2 Zero-Shot Learning for Events and Activities

Another challenge is moving towards identifying more abstract events and activities in videos, robot perception, or other modalities. While it has become fairly easy to recognize various sorts of concrete physical entities, such as *human heads* or *bottles of wine*, more abstract phenomena remain challenging.

Consider, e.g., the concept of *marriage proposals*. They obviously involves humans, but these humans can behave in a range of different ways, and the proposal could occur in almost arbitrary settings. Moreover, in a longer video, the majority of footage could cover the preparations and set-up, while the actual proposal may be rather short.

Many human activities and events – consider also *socializing*, *traveling*, or *earthquakes* – are categories that exhibit enormous variability in their visual appearance. Learning them thus works best when there are many training examples. Unfortunately, the number of possible categories of videos that people upload to platforms like YouTube is unbounded, so ideally we would want systems that can learn with minimal amounts of training data per category.

A recent approach to zero-shot learning for video [Gan et al. 2016] addresses this by jointly exploiting multiple forms of knowledge. Given a new category name q as a textual query, we first determine related concepts of interests. For this, we draw on word vectors trained on a large textual corpora. We then determine which of these concepts can reliably be detected in the videos.

Subsequently, our system uses a sophisticated procedure to aggregate the concept-specific ranking scores, based on the recovery of a low-rank order matrix from multiple pairwise order matrices for different concept ranking lists. This finally results in a ranked list of videos for the input query.

2.3 Towards Visual Semantics

Recently, there has been growing interest in systems that fully bridge visual analytics and natural language processing in tasks such as automatic image and video caption generation [Rohrbach et al. 2015; Venugopalan et al. 2015]. First, one obtains a visual descriptor of the input image or video, e.g. using a deep convolutional neural network. Then a statistical machine translation engine or a recurrent neural network with LSTM or GRU units is used to predict an output string, e.g. using a greedy approach or beam search for decoding.

Recently, the field has also been making initial inroads on addressing visual question answering, a task that has occasionally been touted as a possible multimodal replacement for the Turing test [Marcus 2014]. In fact, it appears that the same family of techniques as for captioning can be used for such question answering [Antol et al. 2015].

Still, while these sorts of algorithms produce remarkable captions or answers in many cases, they remain highly brittle. In some cases, their output is entirely wrong, revealing significant gaps in their understanding of the image. They also support only a limited vocabulary, focusing on lower-level descriptions than humans typically give. For example, a system may describe a video as portraying four humans standing around a table but fail to recognize that this is a birthday party.

Addressing this will require advances at both the algorithmic level and at the data level. In Section 4, we show how one can mine rich knowledge from multiple modalities to address some of these needs.

3. TEXT ANALYTICS BY LEARNING FROM HETEROGENEOUS DATA

When children learn a language, this learning does not happen in a vacuum, but in a rich environment, in which they receive various forms of implicit and at times also explicit feedback. Chomsky famously argued that the linguistic feedback given to children is insufficient for learning a language without prior knowledge. Many others disagree with this assertion. Either way, it is clear that multimodal experience is important for at least some aspects of child language acquisition.

Thus, we expect that multimodal content can be an asset for text analytics as well. In recent years, there has been enormous growth in the amount of available multimodal content, due to the 24/7 ubiquity of mobile devices as well as convenient online sharing platforms [Thomee et al. 2016].

3.1 Cognitive Language Modeling

While computers are getting better and better at solving highly non-trivial tasks, systems such as Google Translate still make silly mistakes that a human is unlikely to make. Given the German sentence “Das Bier bestellte der Vater” (“The father ordered the beer”), it happily returns the odd English translation “The beer ordered the father”. Computers don’t really know that ordering a beer is something a father would do but that ordering a father is not something a beer would do.

In a recent work, we attempted to address this problem by using statistics from multimodal data [Shutova et al. 2015] to better assess which such expressions make sense. Our approach focuses on verb selectional preferences. The standard method is to collect large-scale statistics from a syntactically parsed corpus. However, one observes that certain combinations, while perfectly reasonable from a cognitive perspective, are not salient enough in the data. For instance, one often finds that the verb *to cut* occurs primarily in the financial sense of cutting interest rates or taxes, but not with the word *knife*. Often, the problem is that the original primary meaning of a word is less frequent than more abstract metaphorical uses.

We addressed this by collecting statistics from large amounts of multimodal data, based on the assumption that images and video more likely reveal primary meanings in contrast to the more abstract ones used in text. Specifically, we used the recent YFCC100M dataset, which contains over 99 million images and around 800,000 videos [Thomee et al. 2016].

Our multimodal semantic approach extracts predicate-argument relations from text, images, and videos. For the latter two, it draws on metadata tags rather than grammatically coherent sentences. Still, the roles that individual words play appear discernible from their visual semantic context, as manifested by the other tags in a given set. We use NLP resources such as WordNet and VerbNet as well as integer-linear programming to achieve this.

When we combine the resulting statistics with the standard kind of large-scale statistics from text, we obtain a better indication of whether a given sequence of words is plausible or not.

3.2 Metaphor Interpretation

Cognitive language modeling of the sort mentioned above can also be applied to the task of metaphor interpretation. Metaphor is remarkably ubiquitous in human language. When we are *aiming at reaching a consensus*, we are neither physically aiming a projectile nor arriving at a location.

We can automatically replace metaphoric expressions with literal ones using the same multimodal cognitive language modeling techniques [Shutova et al. 2015]. For example, *a carelessly leaked report* becomes *a carelessly disclosed report*. For this, we choose replacements that are cognitively plausible in the given context, provided that they also share some aspects of meaning with the original metaphorical words being replaced.

4. TOWARDS MULTIMODAL KNOWLEDGE HARVESTING

In order to better address the needs of both visual analytics and text analytics, a promising direction is to move towards increasingly richer data. Rich multi-faceted data has the potential to be re-used across a range of different tasks. This is similar to how humans, rather than starting from scratch, draw on a rich network of prior knowledge and experience when acquiring a new concept or faced with a task.

Over the past years, knowledge graphs describing millions of entities and facts have grown to become standard assets at companies such as Google, Microsoft, Yahoo!, and Baidu. They use them not only for their Web search engines but also in various applications.

Freely available knowledge graphs such as YAGO [Hoffart et al. 2011] played an important role in IBM's Jeopardy!-winning Watson system, especially for filtering answers based on their semantic type. For instance, given a long input question, Watson first primarily relies on semantic similarity to identify candidate answers. Subsequently, if it can recognize that the question is aiming a British philosopher, for instance, the list of candidates can be narrowed down significantly, by retaining only those entities likely to be British philosophers.

These large knowledge graphs, however, focus not on multimodal and commonsense knowledge but on general encyclopedic information, e.g. the fact that the capital of New Jersey is called Trenton and has a population of around 85,000.

4.1 Multimodal and Commonsense Knowledge

The influential ImageNet [Deng et al. 2009] organizes millions of images with respect to the WordNet lexicon, covering a total of 21,841 concepts with an average of over five hundred images per concept.

De Melo & Weikum presented a method to automatically collect Flickr images and YouTube videos for a large number of concepts in WordNet [de Melo and Weikum 2010].

Visual Genome [Krishna et al. 2016] is a recent crowdsourced knowledge base covering currently around 75,000 unique concepts, providing very large amounts of attributes, relationships, and regions in images for them, as well as image question/answer pairs.

The MENTA knowledge graph extended WordNet's taxonomy even further with hundreds of thousands of finer-grained classes, e.g. *Churches in the Netherlands* or *Mercedes-Benz vehicles*, using images and videos hosted on Wikipedia and Wikimedia servers [de Melo and Weikum 2014].

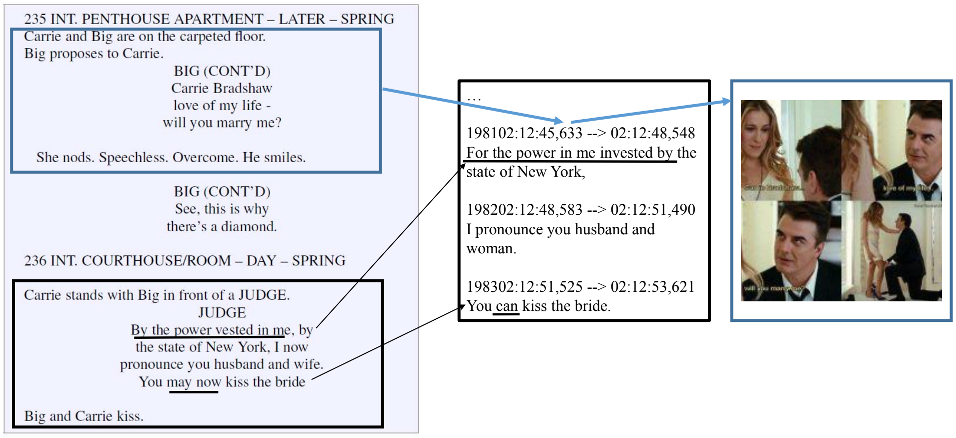
In a series of works, we have been aiming to collect commonsense knowledge about the world. Much of our initial work focused on information extraction from text. We developed new approaches for harvesting knowledge from Web-scale n-grams [Tandon et al. 2011]. In the WebChild project [Tandon et al. 2014], we developed new approaches to disambiguate properties of objects, e.g. *grass hasColor green* and *chili hasTaste spicy*, and mined comparative knowledge [Tandon et al. 2014], e.g. that elephants are bigger than cats. Another recent effort [Tandon et al. 2016] mined part-whole relationships from text and from image tags, while also determining which of these are visible.

4.2 Knowlywood

Our Knowlywood system [Tandon et al. 2015] goes beyond the physical object-focused knowledge of most previous efforts, by harvesting detailed information about higher-level activities, e.g. *delivering a speech* or *announcing an engagement*. Such activities are much harder to recognize in multimodal data and also come in a variety of different forms. While it is possible to overcome this difficulty using zero-shot learning, as mentioned earlier, it is often beneficial to have rich background knowledge about a given activity, as this has the potential to enable higher-quality recognition and additional inference for more detailed interpretations.

To obtain such knowledge about activities, our Knowlywood project taps into Hollywood movie scripts, as well as scripts for TV series, sitcoms, and other narrative text such as novels. Truly understanding what is going on in a movie is still very hard for computers. Working with text, while also challenging, is in some ways easier. Fortunately, most movies come with closed captions or audio descriptions. These in turn, can be linked to movie scripts, which provide even further information. These sources have abundant coverage of everyday activities, albeit still often in implicit form (e.g., in a dialog). Especially scripts have an explicit structure, though, with scenes that start with short descriptions of location, time, and characters involved. Natural language understanding methods can then be applied to these movie scripts to extract information about human activities.

Here, an *activity* is given as a pair (v, o), where v is a disambiguated verb or verb phrase and *o* is a disambiguated noun or noun phrase. For example, Knowlywood describes the activity *climbing a mountain*, providing information about participants (such as a human, or, more specifically, a climber), typical location, and typical time of day. The activities are semantically grouped in terms of synonymy, hierarchically aligned, and come with temporal links to typical preceding or following activities. Finally, activities also come with representative multimodal content.



235 INT. PENTHOUSE APARTMENT – LATER – SPRING
Carrie and Big are on the carpeted floor.
Big proposes to Carrie.
BIG (CONT'D)
Carrie Bradshaw
love of my life -
will you marry me?

She nods. Speechless. Overcome. He smiles.

BIG (CONT'D)
See, this is why
there's a diamond.

236 INT. COURTHOUSE/ROOM – DAY – SPRING

Carrie stands with Big in front of a JUDGE.
JUDGE
By the power vested in me, by
the state of New York, I now
pronounce you husband and wife.
You may now kiss the bride

Big and Carrie kiss.

...

198102:12:45,633 --> 02:12:48,548
For the power in me invested by the
state of New York,

198202:12:48,583 --> 02:12:51,490
I pronounce you husband and
woman.

198302:12:51,525 --> 02:12:53,621
You can kiss the bride.

Fig. 1. Excerpt from a movie script and corresponding multimodal content, aligned via dialogs in the movie script and subtitles

To mine this knowledge, Knowlywood operates in a three stage pipeline. First, it syntactically analyses the input scripts and identifies clauses within sentences, while extracting candidate phrases for activities and computing statistics. This stage already produces large amounts of knowledge about activities, but with substantial noise and many false positives. We then rely on Probabilistic Soft Logic for further inference, subject to soft constraints, to clean the candidate set and construct the final set of high-quality activity knowledge.

Additionally, we attach video frames (including sound snippets) to activities. We record the provenance of each activity in the text. If there is visual data corresponding to the provenance text, then we can attach that data to the activity. For movies, the subtitle data include timestamps that can point us to the movie frames at that timestamp. Scripts contain nearly the same dialogues as the subtitles, thus making an alignment easy. The TV series and sitcom data that we use sometimes contains video frames as part of an episode description. We use the image caption and its position in the surrounding HTML page to heuristically align the image with the text of its corresponding scene.

5. CONCLUSION AND OUTLOOK

The massive availability of images, video, and text on the Web has led to formidable challenges for information management, but also to new opportunities for intelligent systems. In this paper, we have presented examples of systems that attempt to move beyond regular supervised pattern recognition settings, towards models with a more sophisticated understanding of the world. This involves building up knowledge by drawing on multiple modalities and later being able to draw on this prior knowledge.

The next step will be to find ways to feed this information back into the algorithms that drive our AI systems. This is harder than it seems, because the type of knowledge we typically extract from movies, images, and text is not always of the form expected by our most powerful algorithms. One possible avenue is to develop algorithms that jointly learn from multiple heterogeneous sources and produce rich neural representations [Chen et al. 2015]. These encodings can then be exploited in a variety of deep learning algorithms.

In the future, we believe that such multimodal knowledge will enable novel forms of cognitively inspired models that lead to significant progress in getting machines to behave in line with human expectations.

ACKNOWLEDGMENTS

Gerard de Melo's research is supported by China 973 Program Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003, 61550110504.

REFERENCES

ANTOL, S., AGRAWAL, A., LU, J., MITCHELL, M., BATRA, D., ZITNICK, C. L., AND PARIKH, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

CHEN, J., TANDON, N., AND GERARD DE MELO. 2015. Neural word representations from large-scale commonsense knowledge. In *Proceedings of WI 2015*.

DE MELO, G. AND WEIKUM, G. 2010. Providing multilingual, multimodal answers to lexical database queries. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*. ELRA, Paris, France, 348–355.

DE MELO, G. AND WEIKUM, G. 2014. Taxonomic data integration from multilingual Wikipedia editions. *Knowledge and Information Systems 39,* 1 (April), 1–39.

DENG, J., DONG, W., SOCHER, R., LI, L., LI, K., AND LI, F. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 248–255.

GAN, C., LIN, M., YANG, Y., DE MELO, G., AND HAUPTMANN, A. G. 2016. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*. AAAI Press.

GE, T., WANG, Y., DE MELO, G., SHARF, A., AND CHEN, B. 2016. ShapeExplorer: Querying and exploring shapes using visual knowledge. In *Proceedings of EDBT 2016*.

HOFFART, J., SUCHANEK, F. M., BERBERICH, K., LEWIS-KELHAM, E., DE MELO, G., AND WEIKUM, G. 2011. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, Eds. ACM, New York, NY, USA, 229–232.

KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANDITIS, Y., LI, L.-J., SHAMMA, D. A., BERNSTEIN, M., AND FEI-FEI, L. 2016. Visual Genome: Connecting language and vision using crowdsourced dense image annotations.

MARCUS, G. 2014. What Comes After the Turing Test? *The New Yorker, June 9, 2014*.

ROHRBACH, A., ROHRBACH, M., TANDON, N., AND SCHIELE, B. 2015. A dataset for movie description. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

SHUTOVA, E., TANDON, N., AND DE MELO, G. 2015. Perceptually grounded selectional preferences. In *Proceedings of ACL 2015*. 950–960.

TANDON, N., DE MELO, G., DE, A., AND WEIKUM, G. 2015. Knowlywood: Mining activity knowledge from Hollywood narratives. In *Proceedings of CIKM 2015*.

TANDON, N., DE MELO, G., SUCHANEK, F. M., AND WEIKUM, G. 2014. WebChild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of ACM WSDM 2014*. 523–532.

TANDON, N., DE MELO, G., AND WEIKUM, G. 2011. Deriving a Web-scale common sense fact database. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence (AAAI 2011)*. AAAI Press, Palo Alto, CA, USA, 152–157.

TANDON, N., DE MELO, G., AND WEIKUM, G. 2014. Acquiring comparative commonsense knowledge from the web. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*. AAAI, 166–172.

TANDON, N., HARIMAN, C., URBANI, J., ROHRBACH, A., ROHRBACH, M., AND WEIKUM, G. 2016. Commonsense in parts: Mining part-whole relations from the web and image tags. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*.

TAYLOR, A., MARCUS, M., AND SANTORINI, B. 2003. *Treebanks: Building and Using Parsed Corpora*. Springer Netherlands, Dordrecht, Chapter The Penn Treebank: An Overview, 5–22.

THOMEE, B., ELIZALDE, B., SHAMMA, D. A., NI, K., FRIEDLAND, G., POLAND, D., BORTH, D., AND LI, L.-J. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM 59,* 2 (Jan.), 64–73.

VENUGOPALAN, S., ROHRBACH, M., DONAHUE, J., MOONEY, R., DARRELL, T., AND SAENKO, K. 2015. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

VINYALS, O., KAISER, L. U., KOO, T., PETROV, S., SUTSKEVER, I., AND HINTON, G. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2755–2763.

Gerard de Melo is an Assistant Professor at IIIS, Tsinghua University, where he is heading the Web Mining and Language Technology Group. He has published over 50 research papers in these areas, being awarded Best Paper awards at CIKM 2010, ICGL 2008, and the NAACL 2015 Workshop on Vector Space Modeling, as well as an ACL 2014 Best Paper Honorable Mention, a Best Student Paper Award nomination at ESWC 2015, and the WWW 2011 Best Demonstration Award, among others.

Niket Tandon is a final-year doctoral candidate at the Max Planck Institute for Informatics. His thesis is on automated multimodal methods to acquire commonsense knowledge, and he has published extensively on this.