Visual Attributes for Enhanced Human-Machine Communication*

Devi Parikh[1]

***Abstract*— In computer vision systems today, humans typically communicate with the machine via limited interactions e.g. providing coarse image labels. This seems rather wasteful because it is precisely the human abilities that we aim to replicate in automatic image understanding. Moreover, humans are often meant to interact with vision systems as users (e.g. image search) or as supervisors training the system - be it for niche applications or for generic visual concepts such as everyday objects and scenes. On the flip side, machines today also rarely communicate with humans. Vision models are often complex and non-transparent. They simply fail without explaining why which is frustrating for users and perplexing for researchers. Here we describe some of our recent efforts towards using attributes to enhance the mode of communication between humans and machines to improve visual recognition.**

I. Introduction

Building visual recognition systems today typically involves collecting images, having humans label them, extracting features, and training complex machine learning algorithms. The human-provided labels are typically very coarse (e.g., “this is a horse”, “this is not”). While the use of statistical models has contributed significantly to progress in visual recognition in the last two decades, it leaves human knowledge underutilized. Since it is precisely human recognition abilities that we aim to replicate in automatic semantic image understanding, such restrictive human involvement seems wasteful. Computer vision systems can be made more accurate if they tap into the vast, detailed, common-sense knowledge humans have about the visual world.

On the flip side, today’s vision systems typically output equally coarse decisions (e.g., binary labels for images or bounding boxes around objects). They often fail, and do so disgracefully without any warnings or explanations. Even imperfect systems can be useful if they communicate with non-experts, allowing users to understand the system’s capabilities and limitations.

We propose to enrich human-machine communication to improve visual recognition by exploiting visual attributes. Visual attributes are mid-level concepts such as properties of materials (*furry*, *metallic*), spatial layouts (*open*, *congested*), or faces (*young*, *female*) that bridge the gap between low-level image features (e.g. texture) and high-level concepts (e.g. beach, car, Jane Doe). Attributes are shareable across different but related concepts. Most importantly, attributes are both visual (i.e. machine detectable) and semantic (i.e. human understandable), making them ideal as a mode of communication between the two. The use of visual attributes has received a lot of attention in computer vision over the past few years [1–18]. They have been used as better representations for image classification [18], detailed scene parsing [17], generating textual descriptions of images [3], learning models of visual categories from textual descriptions [1] and as keywords for image search [4]. We propose a novel use of attributes: enhancing human-machine communication.

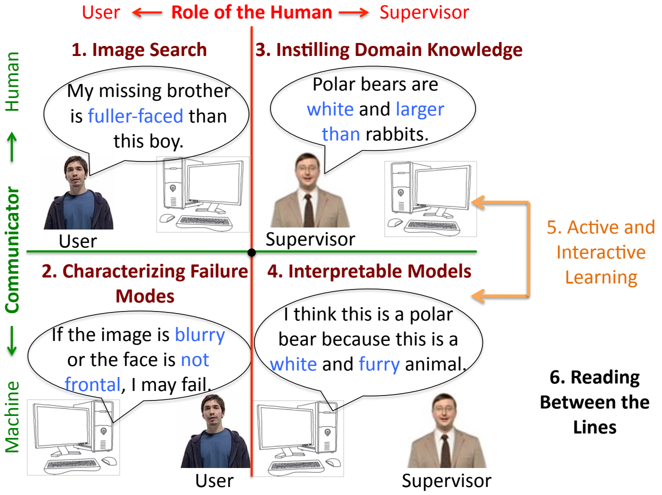
User ← Role of the Human → Supervisor
1. Image Search
3. Instilling Domain Knowledge
Human
Communicator
Machine
My missing brother is fuller-faced than this boy.
User
Polar bears are white and larger than rabbits.
Supervisor
5. Active and Interactive Learning
2. Characterizing Failure Modes
4. Interpretable Models
If the image is blurry or the face is not frontal, I may fail.
I think this is a polar bear because this is a white and furry animal.
6. Reading Between the Lines
User
Supervisor

Fig. 1. Overview of scenarios where we exploit enhanced human-machine communication. We identified two dimensions that allow for a systematic exploration of the space of challenges and opportunities.

Progress in visual recognition hinges on an effective mode of communication between humans and machines due to several factors. (1) The large amount of visual media on the web and proliferation of digital-imaging devices such as cameras, cell phones, webcams and wearable computing such as Google Glass present us with a critical need for technologies that allow consumers to search through, organize and interact with visual media. (2) Applications of vision in biology, astronomy and medical imaging crave automation, but the automation can succeed only if it incorporates the expert’s domain knowledge. (3) There is a significant need for building interpretable algorithms and effective human-machine teams for semi-autonomous systems, both for converting streams of data into actionable information and for building the operator’s trust. There are numerous examples of technologies that outperform humans but go unused because of insufficient user or operator trust [19]. (4) Finally, the advent of crowdsourcing presents us with tremendous opportunities to leverage vast amounts of basic human knowledge to improve computer vision. In all these cases, progress relies critically on having effective means for humans to communicate with machines as well as for

* This is an invited paper associated with an invited talk at the 51st Annual Allerton Conference on Communication, Control, and Computing.

[1] Devi Parikh is with Faculty of the Bradley Department of Electrical and Computed Engineering at Virginia Tech in Blacksburg, VA, USA. parikh@vt.edu

machines to communicate with humans.

II. Applications

We have organized our research agenda along two dimensions resulting in four scenarios (Fig. 1). The first dimension indicates which agent is the communicator conveying the semantics (human or machine), and the second indicates the human's role (user or supervisor). Notice that most visual recognition systems involve the human in at least one of these roles, and often both. Effective attribute-based human-machine communication in these scenarios can lead to (1) Improved image retrieval for applications in law enforcement, missing person search, post-disaster family reunification (user: "My missing son looks like this but thinner.") and personalized on-line shopping (user: "I want shoes shinier than these.") [20]. (2) Self-evaluating and more reliable systems (machine: "If the image is blurry or the face is not frontal, I may fail at recognizing the person."). (3) Opportunities for humans to instill their common sense knowledge about the visual world in machine algorithms [2] by effectively combining traditional AI-like representations with statistical models. (4) Interpretable algorithms (machine: "I think this is a bear because it is a large, brown, furry animal"). (5) Transformative active and interactive learning [21, 22] (machine: "Is this not an alley because it is too natural?") (6) Models of nuances in human behavior that enable reading between the lines of what a user explicitly states resulting in improved machine performance without added human effort [23]. For instance, if a user wants "a white, furry dog", the system should not return images of white, furry dogs that are barking viciously. While they technically match the query, it is unlikely that the user would fail to mention a salient attribute such as vicious if that is what she was looking for [24]. If a user wants "photographs of a couple where the woman is smiling more than the man", the system should not return pictures where the couple is frowning, even if the man is frowning more than the woman. Even though a smiling relative attribute model may score the woman as smiling more than the man (i.e. match the user's query), what is implicit in the user's description is that the man and woman should both be smiling *and* the woman should be smiling more than the man [25].

References

[1] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009.

[2] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011.

[3] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[4] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.

[5] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV*, 2010.

[6] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *BMVC*, 2009.

[7] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *ICCV*, 2009.

[8] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *ECCV*, 2010.

[9] V. Ferrari and A. Zisserman, "Learning visual attributes," in *NIPS*, 2007.

[10] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *ECCV*, 2010.

[11] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba, "Semantic label sharing for learning with many categories," in *ECCV*, 2010.

[12] G. Wang, D. Forsyth, and D. Hoiem, "Comparative object similarity for improved recognition with few or no examples," in *CVPR*, 2010.

[13] D. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *ICCV*, 2011.

[14] A. Kovashka, S. Vijayanarasimhan, and K. Grauman, "Actively selecting annotations among objects and attributes," in *ICCV*, 2011.

[15] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *CVPR*, 2011.

[16] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *ICCV*, 2011.

[17] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *CVPR*, 2010.

[18] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012.

[19] J. Stack, "Automation for underwater mine recognition: Current trends & future strategy," in *Proceedings of SPIE Defense & Security*, 2011.

[20] A. Kovashka, D. Parikh, and K. Grauman, "Whittle-Search: Image search with relative attribute feedback," in *CVPR*, 2012.

[21] A. Parkash and D. Parikh, "Attributes for classifier feedback," in *ECCV*, 2012.

[22] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers & attributes via relative feedback," in *CVPR*, 2013.

[23] D. Parikh and K. Grauman, "Implied feedback: Learning nuances of user behavior in image search," in *ICCV*, 2013.

[24] N. Turakhia and D. Parikh, "Attribute dominance: What pops out?" in *ICCV*, 2013.

[25] A. Sadovnik, A. C. Gallagher, D. Parikh, and T. Chen, "Spoken attributes: Mixing binary and relative attributes to say the right thing," in *ICCV*, 2013.