Computational Visual Recognition

Alexander C. Berg

Computational visual recognition concerns identifying *what* is in an image, video, or other visual data, enabling applications such as measuring location, pose, size, activity, and identity as well as indexing for search by content. Recent progress in economical sensors and improvements in network, storage, and computational power make visual recognition practical and relevant in almost all experimental sciences and in commercial applications from everyday robotics and surveillance to video games and online image search. **My work in visual recognition brings together *large-scale machine learning*, insights from *psychology and physiology*, computer *graphics*, *algorithms*, and a great deal of *computation.***

This setting provides an excellent perspective on general problems of dealing with enormously large datasets consisting of very high dimensional signals with interesting structure—for computer vision, both *structure from the physical world* around us and from *how robots or people reason* about and communicate regarding the world. Our visual environment makes up much of our daily lives, and as sensors and computers become more capable, automating understanding and exploiting visual structure is increasingly relevant both in academia and industry.

My work spans the study of recognition from low-level feature representations for images [15, 12, 4, 2] to modeling [2, 26, 3] with a strong thread of large-scale machine learning for computer vision [13, 29, 23, 22, 10]. I am generally interested in expanding the range of targets for computer vision [17, 19, 20, 6, 3] and closing the loop between what we recognize and how people describe the world using natural language [25, 28, 1, 21, 11, 18, 5].

As long-term goals I want to broadly expand the domain where computational visual recognition is effective, as well as to use it as a tool for allowing automated systems to be aware of the environment around us, making for better *robots*, *smart environments*, and *automated understanding of human language and perception*. Also, despite recent progress in computer vision, there is still no equivalent of the prescriptive decision trees and recipes we have for fields such as numerical analysis. To this end, another long-term goal is to help promote development of prepackaged computer vision techniques and guides for researchers and practitioners in other fields. As just one example, it would be great to allow scientists in microbiology to *interactively* develop a state of the art detector for a cell structure of interest, and then immediately run this detector on millions of archived images! One of the fundamental keys to accomplishing these goals is developing approaches and algorithms to address the challenge of efficiently learning models over enormous and complex data—both from images, video, and 3D structure, as well as associated text and social network data.

The rest of this research statement describes some of my work and some future directions:

- **Large-Scale Recognition** – Studying recognition in the context of millions to billions of images, using complex features, and a large output space (what to recognize). The goals are to understand the structure of the problem, and to develop approaches for making sense of the outputs of recognition at large scale.

- **Large-Scale Machine Learning for Recognition**—The scale of vision data can be immense, beginning with images or video, then considering enormous numbers of sub-regions and large spaces of interdependent labels – configurations of parts, segmentation, etc. Effectively learning models at this scale is challenging.

- **Connections between NLP and Computer Vision**—What people describe about the visual world using language is a useful proxy for what we might want to recognize. By studying the connection between language and computer vision we can (sometimes automatically) identify new targets for visual recognition, as well as automatically collect noisy labels for visual content.

- **Future Directions for Situated Recognition**—Despite great progress in recognition of "internet images" there has been less focus from the core computer vision community on recognition in the everyday world around us. As cameras and devices proliferate, such imagery will scale far beyond what we consider as internet-scale vision today, and involves contextual structure and tasks that are not addressed by current recognition research. One of my main research thrusts going forward is to address this lack, enabling advances in automated, interactive, understanding of our daily world to aid human-computer interaction, human-robot interaction, and robotics.

- **Collaborations**—I am fortunate to have been able to pursue my research interests in collaborations on a wide range of problems from neuroscience to environmental science and nanomaterials research, as well as more traditional applications including image search.

All of this work is part of an attempt to understand the structure of visual data and build better systems for extracting information from visual signals. Such systems are useful in practice because, although for many application areas human perceptual abilities far outstrip the ability of computational systems, automated systems already have the upper hand in running constantly over vast amounts of data, e.g. surveillance systems, process monitoring, or indexing web images, and in making metric decisions about specific quantities such as size, distance, or orientation, where humans have difficulty.

My research activities to date reflect my perspective that theoretical pursuits should be combined with systems building and service to applications in related scientific and commercial communities for a balanced research program in visual recognition. An important part of this is interacting with the vibrant computer vision, machine learning, and related communities through publications, conferences, workshops, working with scientists in fields that can benefit from computer vision, and consulting for industry.

Impact: *This section is to address various aspects of impact for the tenure process at UNC Chapel Hill.* According to Google Scholar, my 54 papers have over 7000 citations (with > 3000 citations since I joined UNC). Sixteen of these papers have over 100 citations, and five have over 500. To a degree this reflects the fact that some of my work has made core contributions to areas of computer vision: local descriptors and image alignment [2], descriptors for motion [12], machine learning for recogniton [22, 23, 29], large-scale computer vision and benchmarking [27], and connections between natural language processing (NLP) and computer vision [18, 21]. Our work using computer vision to study entry-level categories (the language people use to refer to objects in images) won the *2013 Marr Prize* [25].

I have graduated 10 MS students (4 at UNC and 6 at Stony brook). At UNC, my first direct PhD advisee (Xufeng Han) will graduate December 2015, and another will graduate by December 2016. I have 4 more junior PhD student advisees in the pipeline as well as two co-advised students.

I have designed and taught a new undergraduate course on computational photography, addressing both low-level image processing and techniques for manipulating imagery using the Matlab, matrix-centric, scientific computing environment. I will teach this course again in Spring 2016, and it will be added with a permanent course number.

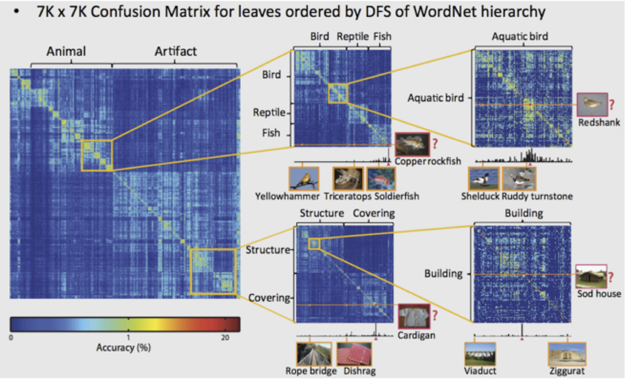
A significant part of my faculty career has been working with wide-ranging set of graduate students. Several of the students with whom I have had close collaborations are now faculty or on the academic job market. Here is a partial list along with the number of papers and current submissions we have co-authored: Subhransu Maji (faculty, UMass Amherst, 4 papers), Jia Deng (faculty, Michigan, 7 papers), Kota Yamaguchi (faculty, Tohoku, 5 papers), Olga Russakovsky (post-doc, CMU, 3 papers), Vicente Ordonez (faculty, University of Virginia, 5 papers).

Funding: In addition to the wonderful support of UNC Chapel Hill, my research is mainly funded by the National Science Foundation (NSF). I am the sole PI on a CAREER grant (2015-2019, $\$509,965$) for situated visual recognition in the world around us as opposed to internet images. I am the UNC PI on two collaborative NSF grants with other institutions: A National Robotics Initiative (NRI) grant on developing visual recognition capabilities for robots (2015-2018, $\$269,480$ UNC), and an exploiting parallism and scalability (XPS) grant on developing an FPGA cloud for machine learning for deep networks (2015-2018 $\$324,707$ UNC) to build vision systems that learn about and recognize the world faster. I am also a co-PI on a UNC-only Cyber Physical Systems (CPS) grant on scheduling for GPU-based computation for autonomous vehicle driving systems (2015-2019, $\$1,000,000$). I have also received faculty gifts from Google ($\$50,000$) to support research on machine learning, from Facebook ($\$25,000$) to support ILSVRC, and a US Air Force research labs supplement ($\$50,000$) for the CPS grant to support developing object labeling in video infrastructure as one step toward improving vision systems for video.

My research activities to date reflect my perspective that theoretical pursuits should be combined with systems building and service to applications in related scientific and commercial communities for a balanced research program in visual recognition. An important part of this is interacting with the vibrant computer vision, machine learning, and related communities through publications, conferences, workshops, working with scientists in fields that can benefit from computer vision, and consulting for industry.

Large Scale Recognition

As computational visual recognition is beginning to work, the field is moving toward both applying recognition to larger datasets as well as recognizing larger and richer label spaces—the possible outputs for recognition. One major question is, "what difference does large scale make?" Some of my work [8, 10] attempts to begin answering this question by analyzing how the difficulty of image recognition problems change as the number of possible labels to output increase. The confusion matrix shows at a glance that not all categories are equally distinguishable. Our work goes on to show a direct correlation between the "denseness" of sets of potential labels in the WordNet hierarchy and the difficulty for algorithms to visually distinguish between images depicting elements in those sets. This correlation was surprising, as WordNet is curated from a linguist's point of view without directly taking into account visual appearance. Ours was the first academic work to consider high level categorization with such a large set (10,000) of categories of images.



Beyond using such large-scale label spaces to study the difficulty of recognition, they can be useful as the basis for high level feature representations. In later work [7] we show that the output of large collections of classifiers make a very strong feature representation for identifying similar images—even when the accuracy of the individual classifiers is quite low. Such methods can be used to identify images similar to out of band queries for which no examples are available in training. Furthermore we show how to exploit hierarchical structure in the label space to improve similarity. This work significantly improved on the previous state of the art (from Google Research) for large scale similar image retrieval in some settings. Our results represent a departure from previous similarity learning techniques that focus on directly learning similarity functions, instead we suggest first learning to recognize somewhat higher level concepts (as the availability of labeled data allows) before attacking similarity. As recognition advances it is unrealistic to "start from scratch" for every new target. This work can be seen as one way to harness the output of previously trained models as features. Other options are discussed later along with connections to natural language.

Moving from classifying images to detecting objects—putting a bounding box around the extent of the object in an image—the effects of scaling up the label space can be challenging. With 10,000 classes of objects, even if each detector has a per-window false positive rate around 1 in a million (roughly the state of the art for some object categories), then an individual detector can be expected to find one false positive every few images, multiplying this by 10,000 results in dozens of false positives per image. How is it possible to make sense of such output? Some of my recent work begins addressing this issue. One possibility I have explored is based on "hedging your bets"—explicitly trading off between the expected accuracy of a prediction and the informativeness of that prediction. For instance a system that is unsure whether a detection is a dog or a cat might hedge its bets and output mammal. In [9] we present an algorithm for optimally tuning a system to produce the most informative predictions (being as specific as possible given the output of noisy classifiers) given a constraint on expected accuracy. This was combined with an object detection framework to make the first demonstration of object *detection* for 10,000 categories of objects exhibited at CVPR 2012 and NIPS 2012.

The future of large-scale recognition is probably to be found in increasing the structure of the output space—for example, recognizing not just that there is a car in the image, but localizing the parts of the car, recognizing attributes of the parts, and the relationship between the car and its surroundings. Early efforts at this type of integration assume relatively simply factored models – *e.g.*, attributes may be assumed to be independent given the part localizations—but joint modeling of complex structured label spaces is currently a challenging open problem.

Large-Scale Machine Learning for Recognition

An intrinsic consequence of large scale recognition problems is the computational challenge of learning models. This is especially true for the complex structured outputs mentioned above. Efficient algorithms are absolutely necessary in order to address the questions we want to ask and this has been a constant aspect of my research.

My work on detection began with the observation that a type of high quality, but slow, classifier[1] used in computer vision (but not for detection) could in fact be evaluated *exponentially* faster than the state of the art [23]. This allowed the classifier to be applied to detection and resulted in improvements in the state of the art for pedestrian detection. In later work we developed efficient training techniques that bring training time down from hours to seconds [22]. This allows approaches that are commonly used in the computer vision community to be applied at much large scale. Our technique has been widely used for large scale classification and detection.

Labeling problems in large scale web corpora can involve 10s of thousands of categories. One way to efficiently deal with large numbers of categories is to use a tree structure over the labels. Our work has demonstrated that is it possible to learn such tree structures efficiently [10]. Also, at this large scale for the label space, nearest neighbor based

[1] based on SVMs with *additive* kernels.

approaches potentially offer advantages in efficiency. Some of my early—and now well cited—work has successfully combined these with the generalization performance of support vector machines [29]. This is related to exciting machine learning work on indefinite kernels.

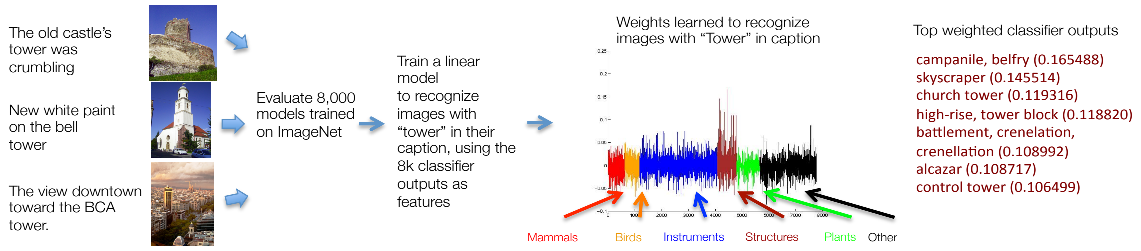
Following up on the similar image retrieval work mentioned in the previous section, we developed a technique for a hash based indexing scheme using the thousands of classifier outputs as features. This provides fast nearest neighbor searches with respect to the learned hierarchical similarity, allowing scaling to very large datasets[7].

I am also very enthusiastic about some of our *negative results* from experiments on large scale classification on ImageNet—we have observed that standard multiclass classification techniques do not scale well in terms of accuracy or efficiency for very large numbers of categories [8]. As a result I have been developing new techniques to use efficient distributed convex optimization to bridge this gap. Our work was the first to distribute training for multiclass models that explicitly optimize multiclass loss (so called single machine methods as opposed to more typical one-vs-the-rest loss) and shows significant advantages in the accuracy vs training time trade-off [13]. The technique is based on a modification of Cramer and Singer's sequential dual method for optimizing a multiclass training objective, augmented by a consensus term based on the alternating direction method of multipliers. In order to address the complex structured prediction problems we expect to become the focus of much recognition work in computer vision in the future, we are developing a more flexible and online, averaged, stochastic gradient descent version of this optimization. This approach works well with the latest deep-learning-based convolutional network features, and can be used to fine-tune such features as necessary.

One of my longer term goals is developing tools to allow *interactive design of high quality object detectors* – bringing state of the art computer vision techniques to a broader set of users in science and industry. Currently too much expertise and time is required to develop detectors, presenting a significant barrier to use. Reducing the design and development time with more efficient learning techniques (to automatically explore the design space) is an important step toward this goal. Our work in progress can simultaneously train high quality detectors for all 20 detectors for the benchmark Pascal VOC detection challenge in less than 2 hours on a single computer without using GPU, an order of magnitude faster than other techniques. Our goal is to bring this down to minutes by developing a combination of more efficient online optimization techniques and highly parallelized feature computation kernels—fast enough to allow interactive exploration of the design space for object detectors.

Connections between NLP and Computer Vision

As computational visual recognition begins to work we are encountering the problem of determining what to recognize! One way to address this problem is by looking at what language people use to describe the visual world. As an example, some of my work tries to expand the space of outputs for visual recognition, *e.g.*, by recognizing attributes of faces [19, 20] and then using these for subsequent recognition tasks (and in the process significantly improving on the then state of the art for face verification). But like much work in recognition, the set of output labels—the facial attributes—were chosen *ad hoc*. In more recent work [6] we exploit text that co-occurs with images as a source of potential attributes for the image content. We then train models using the text as a noisy source of labels and cull models that have poor performance. This represented a first fully automated procedure for discovering a *vocabulary* of attributes and automatically training visual models for those attributes, albeit for shopping images instead of faces


The old castle's tower was crumbling
New white paint on the bell tower
The view downtown toward the BCA tower.
Evaluate 8,000 models trained on ImageNet
Train a linear model to recognize images with "tower" in their caption, using the 8k classifier outputs as features
Weights learned to recognize images with "Tower" in caption
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
0
1000
2000
3000
4000
5000
6000
7000
8000
Mammals
Birds
Instruments
Structures
Plants
Other
Top weighted classifier outputs
campanile, belfry (0.165488)
skyscraper (0.145514)
church tower (0.119316)
high-rise, tower block (0.118820)
battlement, crenelation, crenellation (0.108992)
alcazar (0.108717)
control tower (0.106499)

Toward furthering this connection, I co-organized a 6 week NSF co-sponsored summer workshop for 10 researchers at Johns Hopkins University's Center for Language and Speech Processing in the summer of 2011. Some results were initial studies of what image content people are likely to describe [1] using language as well as text only models that could predict with high accuracy which noun phrases in a descriptive sentence were likely to be depicted in an associated image [11]. Combining observations from these studies with the output of large scale object detections and an optimization based surface realization procedure we also explored automatically generating natural language descriptions of images [21].

Some of our work on building computational models for how people name objects—the idea of entry-level categories from Psychology—was awarded the Marr Prize (best paper award at ICCV) in 2013 [25]. This paper represented some progress toward addressing what we should output from visual recognition systems, as well as continuing to couple research on recognition and natural language processing.

Several of my vision+NLP projects are related to the exciting direction of using computer vision to "ground" language. The figure shows an example from some work in progress, where a classifier is learned to distinguish images with "tower" in their caption from images without "tower" in their caption. The features used are the outputs of a large scale vision system. The features with high weights nicely indicate the meaning of "tower". We can reverse the process to build models of what language people use to describe certain visual content—providing a large-scale computational approach to identifying the "entry-level categories" as we did in [25].

Future work on the connection between NLP and computer vision will likely delve deeper into the syntactic structure of natural language in order to more precisely connect natural language descriptions and visual content (See some intermediate results in out upcoming paper [28]). In addition as the target space for visual recognition becomes larger and more complex, jointly modeling targets becomes critical. Here language can provide estimates of the correlation structure between labels that may allow joint modeling at a larger scale.

Future Directions for Situated Recognition

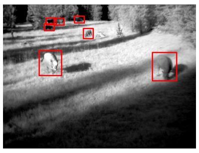
As one future direction of research I propose a new focus for recognition in computer vision, ***situated recognition***. The goal is to recognize the local visual world around us every day as we interact with it. Success will allow automated systems to better understand and monitor our daily environment and improve human-computer and human-robot interaction. This research direction is different than the majority of work in recognition that has focused on *internet images* collected from the web. The biases of such web-collected data (often found using text-based search) may lead to models that do not generalize to a particular environment. *Situated recognition* allows exploiting local context, including human interaction and spoken language, to build models specific to an environment and furthermore to the parts of an environment that are important to people.

Looking forward to the near future, when many aspects of our lives will be continuously observed by multiple imaging systems, advancing computer vision in order to extract useful information about our everyday surroundings and activities from the such imagery will become a central challenge. Already surveillance cameras cover many of our daily spaces, cellphones and tablets are present wherever there are people, and wearable devices with cameras are just beginning exponential growth in numbers. At the same time, there is continual improvement in system power efficiency allowing mobile cameras to be on more of the time (e.g., in cell phones that are aware of their user's face pose and gaze direction). This proposal attempts to put progress toward recognition in the individualized contexts of our everyday lives on the front burner. One major possible direction for attacking this problem is estimating the 3D structure of the world from long-range depth sensors like the Velodyne lidar, active RGBD sensors such as Kinect, and alignment and reconstruction from multiple images. I am interested in *an effort orthogonal to work in 3D localization and reconstruction* focusing instead on *recognizing* content both with and without 3D information. In addition, some situated recognition approaches may use efficient recognition systems to selectively target computationally expensive 3D reconstruction efforts in order to conserve power.

The choice of the term ***situated recognition*** has multiple purposes. The first is to emphasize recognition in a particular context, be it a location such as *my office*, or a type of setting, e.g., *in the car*. The second is to connect to work on situated natural language processing (NLP) that emphasizes the need to use "domain-specific information about the non-linguistic situational context of users' interactions," [e.g. Fleischman and Roy CoNLL 2005] applied to both language and vision. In a particular context, if someone verbally indicates the, "keys over there on the counter," it may help identify a confusing region of pixels as being the keys! At the same time, answering the question "where are my keys?" requires something more specific than a generic key detector. This direction is the core of my recently funded NSF Career award and closely related to my recently funded joint grant on perception for robotics.

Collaborations

One of the advantages of working in such a vibrant field is the opportunity for fruitful collaboration. Currently I am working with researchers at the University of North Carolina Chapel Hill, Stony Brook, Stanford, U.C. Berkeley, UMass Amherst, Brookhaven National Laboratory, and Microsoft. Collaborations work especially well when each party brings unique capabilities to the equation. I have had a great long term collaboration with Fei-Fei Li's group at Stanford, where I brought expertise in discriminative visual recognition and large-scale machine learning to complement their extensive background in high-level vision and dataset collection. This has not only resulted in a series of publications, but I had the opportunity to lead the development of the Large Scale Visual Recognition Challenge (LSVRC) part of the ImageNet project and (in the past) of Pascal VOC. Currently I have three grants funding collaborative activities, and NSF NRI grant with Jana Kosecka at George Mason University on computer vision for robotics, a NSF XPS grant with Michael Ferdman and Peter Milder at Stony Brook University on flexible FPGA-based hardware solutions for deep learning, and support from Facebook and the US Air Force for the benchmarking collaboration on ILSVRC. I have also collaborated with researchers outside of computer science, *e.g.*, environmental scientists working on animal surveillance [24], and material scientists analyzing x-ray scattering images [16]. Currently I have a significant collaboration with Hoi-Chung Leung's lab at Stony Brook on using machine learning to decode neural representations measured by fMRI during short term visual memory [14] and am working to begin collaborations with neuroscientists at other institutions. I have also consulted in industry (including Microsoft Research) – something I feel is an important part of keeping academic research relevant.



Cow surveillance

As ever, I cannot thank my many collaborators and co-authors enough!

Google Scholar: http://scholar.google.com/citations?user=jjEht8wAAAAJ&pagesize=100

[1] A.C. Berg, T.L. Berg, H. Daume III, J. Dodge, A. Goyal, X. Han, A. Mensch, , M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[2] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[3] A.C. Berg, F. Grabler, and J. Malik. Parsing images of architectural scenes. In *Proc. 11th IEEE International Conf. on Computer Vision*, 2007.

[4] A.C. Berg and J. Malik. Geometric blur for template matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[5] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[6] T.L. Berg, A.C. Berg, and J. Shih. Automatic attribute discovery and characterization. In *11th European Conference on Computer Vision*, 2010.

[7] J. Deng, A.C. Berg, and L. Fei-Fei. Hierarchical semantic indexing for large scale retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[8] J. Deng, A.C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *11th European Conference on Computer Vision*, 2010.

[9] J. Deng, J. Krause, A.C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[10] J. Deng, S. Satheesh, A.C. Berg, and L. Fei-Fei. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Neural Information Processing Systems*, 2011.

[11] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daume III, A.C. Berg, and T.L. Berg. Detecting visual text. In *NAACL*, 2012.

[12] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. 9th IEEE International Conf. on Computer Vision*, volume 2, 2003.

[13] X. Han and A.C. Berg. DCMSVM: Distributed parallel training for single-machine multiclass classifiers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[14] X. Han, A.C. Berg, H. Oh, D. Samaras, and H-C Leung. Multiple-voxel pattern analysis of selective representation of visual working memory. in submission.

[15] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] H. Kiapour, K. Yager, A. C. Berg, and T. L. Berg. Materials discovery: Fine-grained classification of x-ray scattering images. In *WACV*, 2014.

[17] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *Computer Vision–ECCV 2014*, pages 472–488. Springer, 2014.

[18] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg. Babytalk: Understanding and generating simple image descriptions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[19] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *Proc. 12th IEEE International Conf. on Computer Vision*, 2009.

[20] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Describable visual attributes for face verification and image search. *Transactions on Pattern Analysis and Machine Intelligence*, Oct 2011.

[21] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012.

[22] S. Maji and A.C. Berg. Max-margin additive models for detection. In *Proc. 12th IEEE International Conf. on Computer Vision*, 2009.

[23] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[24] S. McIlroy, B. Allen-Diaz, and A.C. Berg. Using digital photography to examine grazing in montane meadows. *Journal of Rangeland Ecology & Management*, March 2011.

[25] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Proc. 14th IEEE International Conf. on Computer Vision*, 2013.

[26] X. Ren, A.C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. 10th IEEE International Conf. on Computer Vision*, volume 1, 2005.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[28] Licheng Yu, Eunbyung Park, and Alexander C. Berg andTamara L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[29] H. Zhang, A.C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.