Crowdsourcing in Computer Vision

Adriana Kovashka
University of Pittsburgh
kovashka@cs.pitt.edu

Olga Russakovsky
Carnegie Mellon University
olgarus@cmu.edu

Li Fei-Fei
Stanford University
feifeili@cs.stanford.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. Kovashka, O. Russakovsky, L. Fei-Fei and K. Grauman. *Crowdsourcing in Computer Vision*. Foundations and Trends® in Computer Graphics and Vision, vol. 10, no. 3, pp. 177–243, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-213-6

Foundations and Trends® in Computer Graphics and Vision
Volume 10, Issue 3, 2014

Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2014, Volume 10, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 10, No. 3 (2014) 177–243

DOI: 10.1561/0600000071

Crowdsourcing in Computer Vision

Adriana Kovashka
University of Pittsburgh
kovashka@cs.pitt.edu

Olga Russakovsky
Carnegie Mellon University
olgarus@cmu.edu

Li Fei-Fei
Stanford University
feifeili@cs.stanford.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

Contents

Abstract

Computer vision systems require large amounts of manually annotated data to properly learn challenging visual concepts. Crowdsourcing platforms offer an inexpensive method to capture human knowledge and understanding, for a vast number of visual perception tasks. In this survey, we describe the types of annotations computer vision researchers have collected using crowdsourcing, and how they have ensured that this data is of high quality while annotation effort is minimized. We begin by discussing data collection on both classic (e.g., object recognition) and recent (e.g., visual story-telling) vision tasks. We then summarize key design decisions for creating effective data collection interfaces and workflows, and present strategies for intelligently selecting the most important data instances to annotate. Finally, we conclude with some thoughts on the future of crowdsourcing in computer vision.

A. Kovashka, O. Russakovsky, L. Fei-Fei and K. Grauman. *Crowdsourcing in Computer Vision*. Foundations and Trends® in Computer Graphics and Vision, vol. 10, no. 3, pp. 177–243, 2014.

DOI: 10.1561/0600000071.

1

Introduction

Data has played a critical role in all major advancements of artificial intelligence for the past several decades. In computer vision, annotated benchmark datasets serve multiple purposes:

- to focus the efforts of the community on the next concrete stepping stone towards developing visual intelligence;
- to evaluate progress and quantitatively analyze the relative merits of different algorithms;
- to provide training data for learning statistical properties of the visual world.

We rely on *big data* to move computer vision forward; in fact, we rely on big *manually labeled* data. Harnessing this large-scale labeled visual data is challenging and expensive, requiring the development of new innovative techniques for data collection and annotation. This paper serves to summarize the key advances in this field.

In collecting large-scale labeled datasets for advancing computer vision, the key question is **what** annotations should be collected. This includes decisions about:

- the type of media: simple object-centric images, complex scene images, videos, or visual cartoons;
- the type of annotations: single image-level label, detailed pixel-level annotations, or temporal annotations;
- the scale of annotation: more images with sparse labels or fewer images with more detailed labels.

Different types of data come with different associated costs, including computer vision researcher time (formulating the desired dataset), crowdsourcing researcher time (user interface design and developing the annotation procedure) and annotator time (e.g., finding the visual media to annotate, or providing the semantic labels). There are tradeoffs to be made between the cost of data collection and the resulting benefits to the computer vision community.

There are two ways to optimize this tradeoff between data collection cost and the benefits for the community. The first way is to carefully considering **how** data should be collected and annotated. In some cases annotators may not require any prior knowledge and this effort can be outsourced to an online marketplace such as Amazon Mechanical Turk[1]. As many other crowdsourcing platforms, Mechanical Turk allows "requesters" to post small tasks to non-expert "workers," for low cost per task. The overall cost can still be significant for large-scale data annotation efforts. This can be partially remedied by developing improved user interfaces and advanced crowd engineering techniques.

The second way to optimize the cost-to-benefit tradeoff is directly using existing computer vision algorithms to select **which** data should be annotated. Using algorithms in the loop allows the annotation effort to focus specifically on scenarios which are challenging for current algorithms, alleviating human effort.

The rest of the survey is organized according to these three main questions: what, how, and which data should be annotated. Section 2 discusses key data collection efforts, focusing on the tradeoffs that have been been made in deciding **what** annotations should be collected.

[1] `http://www.mturk.com`

Section 3 dives into the details of **how** to most effectively collect the desired annotations. Section 4 considers the question of **which** data should be annotated and how data collection can be directly integrated with algorithmic development.

The goal of this survey is to provide an overview of how crowdsourcing has been used in computer vision, and to enable a computer vision researcher who has previously not collected non-expert data to devise a data collection strategy. This survey can also help researchers who focus broadly on crowdsourcing to examine how the latter has been applied in computer vision, and to improve the methods that computer vision researchers have employed in ensuring the quality and expedience of data collection. We assume that any reader has already seen at least one crowdsourced micro-task (e.g., on Amazon Mechanical Turk), and that they have a general understanding of the goals of artificial intelligence and computer vision in particular.

We note that most data collection on Mechanical Turk and similar platforms has involved low payment (on the order of cents) for the annotators, and relatively small and often simple tasks (which require minutes to complete), so this is the type of annotation scenario that we ask the reader to imagine. However, crowdsourcing can also involve long-term and more complex interactions between the requesters and providers of the annotation work.

Crowdsourcing is a fairly recent phenomenon, so we focus on research in the past 5-10 years. Some of the most interesting approaches we overview involve accounting for subjective annotator judgements (Sections 2.1.5 and 2.3.2), collecting labels on visual abstractions (Section 2.2.3), capturing what visual content annotators perceive to be similar (Section 2.3.3), translating between annotations of different types (Section 2.4), grouping the labeling of many instances (Section 3.2.1), phrasing data collection as a game (Section 3.2.2), and interactively reducing the annotation effort (Section 4.1, 4.2.1). The contributions we present are both algorithmic, in terms of novel mathematical formulations of solutions to vision problems interlaced with a human annotation effort, and design-based, in terms of accounting for human factors in the implementation and presentation of annotation requests.

References

E. Ahmed, S. Maji, G. Shakhnarovich, and L. S. Davis. Using human knowledge to judge part goodness: Interactive part selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop: Computer Vision and Human Computation*, 2014.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

R. Anirudh and P. Turaga. Interactively test driving an object detector: Estimating performance on unlabeled data. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

S. Antol, C. L. Zitnick, and D. Parikh. Zero-shot learning via visual abstraction. In *European Conference on Computer Vision (ECCV)*. 2014.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5), 2011.

H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *European Conference on Computer Vision (ECCV)*, 2012.

S. Bandla and K. Grauman. Active learning of an action detector from untrimmed videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision (ECCV)*, 2016.

S. Bell, P. Upchurch, N. Snavely, and K. Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (SIGGRAPH)*, 32(4), 2013.

S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transaction on Graphics (SIGGRAPH)*, 33(4), 2014.

S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

A. Biswas and D. Jacobs. Active image clustering: Seeking constraints from humans to complement algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

A. Biswas and D. Parikh. Simultaneous active learning of classifiers and attributes via relative feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *European Conference on Computer Vision (ECCV)*, 2002.

L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

A. Boyko and T. Funkhouser. Cheaper by the dozen: Group annotation of 3D data. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2014.

S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *European Conference on Computer Vision (ECCV)*, 2010.

S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

S. Branson, K. Eldjarn Hjorleifsson, and P. Perona. Active annotation translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. Chandrasekaran, A. Kalyan, S. Antol, M. Bansal, D. Batra, C. L. Zitnick, and D. Parikh. We are humor beings: Understanding and predicting visual humor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

C.-Y. Chen and K. Grauman. Predicting the location of "interactees" in novel human-object interactions. In *Asian Conference on Computer Vision (ACCV)*, 2014.

C.-Y. Chen and K. Grauman. Subjects and their objects: Localizing interactees for a person-centric view of importance. *Computing Research Repository (CoRR)*, abs/1604.04842, 2016.

X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowdsourcing taxonomy creation. In *SIGCHI Conference on Human Factors in Computing Systems*, 2013.

G. Christie, A. Parkash, U. Krothapalli, and D. Parikh. Predicting user annoyance using visual attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. Criminisi. Microsoft Research Cambridge (MSRC) object recognition image database (version 2.0). `http://research.microsoft.com/vision/cambridge/recognition`, 2004.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

J. Deng, O. Russakovsky, J. Krause, M. Bernstein, A. C. Berg, and L. Fei-Fei. Scalable multi-label annotation. In *SIGCHI Conference on Human Factors in Computing Systems*, 2014.

J. Deng, J. Krause, M. Stark, and L. Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(4), April 2016.

A. Deza and D. Parikh. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

I. Endres, A. Farhadi, D. Hoiem, and D. A. Forsyth. The benefits and challenges of collecting richer object annotations. In *IEEE Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2), June 2010.

M. Everingham, , S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge - a Retrospective. *International Journal of Computer Vision (IJCV)*, 2014.

S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, C. Y.-C. Tan, and R. Wang. An automated estimator of image visual realism based on human cognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by Their Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

A. Fathi, M. F. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *British Machine Vision Conference (BMVC)*, 2011.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few examples: an incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*. 2014.

E. Gavves, T. Mensink, T. Tommasi, C. G. M. Snoek, and T. Tuytelaars. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12), 2015.

A. Gilbert and R. Bowden. igroup: Weakly supervised image and video grouping. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

R. G. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015.

S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

Y. Guo and R. Greiner. Optimistic active-learning using mutual information. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31 (10), Oct 2009.

D. Gurari, D. H. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. L. Walker, C. Zhang, J. Y. Wong, and M. Betke. How to collect segmentations for biomedical images? a benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

D. Gurari, S. D. Jain, M. Betke, and K. Grauman. Pull the plug? predicting if computers or humans should segment images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

T. S. Haines and T. Xiang. Active learning using dirichlet processes for rare class discovery and classification. In *British Machine Vision Conference (BMVC)*, 2011.

B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

G. Hua, C. Long, M. Yang, and Y. Gao. Collaborative active learning of a kernel machine ensemble for recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

P. Jain and A. Kapoor. Active learning for large multi-class problems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

P. Jain, S. Vijayanarasimhan, and K. Grauman. Hashing hyperplane queries to near points with applications to large-scale active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

S. D. Jain and K. Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

S. D. Jain and K. Grauman. Active image segmentation propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

K. G. Jamieson, L. Jain, C. Fernandez, N. J. Glattard, and R. Nowak. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

J. H. Janssens. Ranking images on semantic attributes using human computation. In *Advances in Neural Information Processing Systems (NIPS) Workshops*, 2010.

E. Johns, O. Mac Aodha, and G. J. Brostow. Becoming the expert - interactive multi-class machine teaching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods*, 10(1), 2013.

C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 88(2), 2010.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

K. Konyushkova, R. Sznitman, and P. Fua. Introducing geometry in active learning for image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *IEEE International Conference on Computer Vision (ICCV)*, 2013a.

A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *IEEE International Conference on Computer Vision (ICCV)*, 2013b.

A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision (IJCV)*, 114(1), 2015.

A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision (IJCV)*, 115(2), 2015.

A. Kowdle, Y.-J. Chang, A. Gallagher, and T. Chen. Active learning for piecewise planar 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

J. Krause, J. Deng, M. Stark, and L. Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop: Fine-Grained Visual Categorization*, 2013.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2016.

A. Krizhevsky. Learning multiple layers of features from tiny images. `https://www.cs.toronto.edu/~kriz/cifar.html`, 2009.

A. Kulkarni, M. Can, and B. Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*, 2012.

S. Lad and D. Parikh. Interactively Guiding Semi-Supervised Clustering via Attribute-based Explanations. In *European Conference on Computer Vision (ECCV)*, 2014.

C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

D. Larlus, F. Perronnin, P. Kompalli, and V. Mishra. Generating gold questions for difficult visual recognition tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop: Computer Vision and Human Computation*, 2014.

W. Lasecki, M. Gordon, D. Koutra, M. Jung, S. Dow, and J. Bigham. Glance: Rapidly coding behavioral video with the crowd. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2014.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial Pyramid Matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

D. Le, R. Bernardi, and J. Uijlings. TUHOI: Trento Universal Human Object Interaction Dataset. In *Conference on Computational Linguistics (COLING) Workshop: Vision and Language*, 2014.

D.-T. Le, J. Uijlings, and R. Bernardi. Exploiting language models for visual recognition. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

S. Lee and D. Crandall. Learning to identify local flora with human feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop: Computer Vision and Human Computation*, 2014.

F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

X. Li and Y. Guo. Adaptive active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

X. Li and Y. Guo. Multi-level adaptive active learning for scene classification. In *European Conference on Computer Vision (ECCV)*. 2014.

L. Liang and K. Grauman. Beyond comparing image pairs: Setwise active learning for relative attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollÃąr, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.

G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: Human computation algorithms on mechanical turk. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2010.

J. Little, A. Abrams, and R. Pless. Tools for richer crowd source image annotations. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2012.

C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12), 2011.

W. Liu, O. Russakovsky, J. Deng, L. Fei-Fei, and A. Berg. ImageNet Large Scale Visual Recognition Challenge – object detection from video track. `http://image-net.org/challenges/LSVRC/2015/`, 2015.

C. Long and G. Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

C. Long, G. Hua, and A. Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong. Stream-based joint exploration-exploitation active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

S. Maji. Discovering a lexicon of parts and attributes. In *European Conference on Computer Vision (ECCV) Workshops*, 2012.

S. Maji and G. Shakhnarovich. Part discovery from partial correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

S. Maji, L. Bourdev, and J. Malik. Action recognition using a distributed representation of pose and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

A. Mao, E. Kamar, and E. Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.

T. Matera, J. Jakes, M. Cheng, and S. Belongie. A user friendly crowdsourcing task manager. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop: Computer Vision and Human Computation*, 2014.

T. Mensink, J. Verbeek, and G. Csurka. Learning structured prediction models for interactive image labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

A. Montagnini, M. Bicego, and M. Cristani. Tell me what you like and I'll tell you what you are: discriminating visual preferences on flickr data. *analysis*, 10, 2012.

R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

N. S. Nagaraja, F. R. Schmidt, and T. Brox. Video segmentation with just a few strokes. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 2001.

D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011a.

D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, 2011b.

A. Parkash and D. Parikh. Attributes for classifier feedback. In *European Conference on Computer Vision (ECCV)*. Springer, 2012.

G. Patterson and J. Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

G. Patterson, G. V. Horn, S. Belongie, P. Perona, and J. Hays. Tropel: Crowdsourcing detectors with minimal training. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015.

K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

M. Rubinstein, C. Liu, and W. Freeman. Annotation propagation: Automatic annotation of large image databases via dense image correspondence. In *ECCV*, 2012.

M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *European Conference of Computer Vision (ECCV) Workshop: Parts and Attributes*, 2010.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015a. .

O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015b.

B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision (IJCV)*, 2007.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Conference on Learning Theory (COLT) Workshops*. ACM, 1992.

N. B. Shah and D. Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8), 2009.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

B. Siddiquie and A. Gupta. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

G. A. Sigurdsson, O. Russakovsky, A. Farhadi, I. Laptev, and A. Gupta. Much ado about time: Exhaustive annotation of temporal data. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016a.

G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference of Computer Vision (ECCV)*, 2016b.

E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2008.

H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI Conference on Artificial Intelligence Workshop: Human Computation (HCOMP)*, 2012.

O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *International Machine Learning Conference (ICML)*, 2011.

M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

T. Tian and J. Zhu. Max-margin majority voting for learning from crowds. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2, 2002.

D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label MRF optimization. In *British Machine Vision Conference (BMVC)*, 2010.

R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

A. Vezhnevets, J. M. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

S. Vijayanarasimhan and K. Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision (IJCV)*, 91(1), 2011a.

S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011b.

S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *European Conference on Computer Vision (ECCV)*, 2012.

S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision (IJCV)*, 108(1-2), 2014.

S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

S. Vijayanarasimhan, P. Jain, and K. Grauman. Hashing hyperplane queries to near points with applications to large-scale active learning. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(2), 2014.

S. Vittayakorn and J. Hays. Quality assessment for crowdsourced object annotations. In *British Machine Vision Conference (BMVC)*, 2011.

L. von Ahn and L. Dabbish. Esp: Labeling images with a computer game. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, 2005.

L. von Ahn, M. Kedia, and M. Blum. Verbosity: A game for collecting common-sense facts. In *SIGCHI Conference on Human Factors in Computing Systems*, 2006a.

L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *SIGCHI Conference on Human Factors in Computing Systems*, 2006b.

C. Vondrick and D. Ramanan. Video Annotation and Tracking with Active Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision (IJCV)*, 101(1), 2013.

C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting the motivations behind actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

C. Wah and S. Belongie. Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2011a.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011b.

C. Wah, G. Van Horn, S. Branson, S. Maji, P. Perona, and S. Belongie. Similarity comparisons for interactive fine-grained categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

C. Wah, S. Maji, and S. Belongie. Learning localized perceptual similarity metrics for interactive categorization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

D. Wang, C. Yan, S. Shan, and X. Chen. Active learning for interactive segmentation with expected confidence change. In *Asian Conference on Computer Vision (ACCV)*. 2012.

J. Wang, P. G. Ipeirotis, and F. Provost. Quality-based pricing for crowdsourced workers. *NYU-CBA Working Paper CBA-13-06*, 2013.

Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision (ECCV)*. Springer, 2010.

D. S. Weld, Mausam, and P. Dai. Human intelligence needs artificial intelligence. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2011.

P. Welinder and P. Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010.

P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

M. Wigness, B. A. Draper, and J. R. Beveridge. Efficient label collection for unlabeled image datasets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

M. J. Wilber, I. S. Kwak, and S. J. Belongie. Cost-effective hits for relative similarity comparisons. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2014.

M. J. Wilber, I. S. Kwak, D. Kriegman, and S. Belongie. Learning concept embeddings with combined human-machine expertise. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

A. G. Wilson, C. Dann, C. Lucas, and E. P. Xing. The human kernel. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*. 2015.

J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from Abbey to Zoo. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

C. Xiong, D. M. Johnson, and J. J. Corso. Spectral active clustering via purification of the k-nearest neighbor graph. In *Proceedings of European Conference on Data Mining*, 2012.

A. Yao, J. Gall, C. Leistner, and L. V. Gool. Interactive object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

B. Yao, X. Yang, and S.-C. Zhu. Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks. In *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2007.

B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM International Conference on Multimedia*, 2015.

S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *Computing Research Repository (CoRR)*, abs/1507.05738, 2015.

A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

J. Yuen, B. Russell, C. Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

C. Zhang and K. Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.