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
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.
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

1 Introduction 2

2 What annotations to collect 5
   2.1 Visual building blocks ............................... 6
   2.2 Actions and interactions ............................ 15
   2.3 Visual story-telling .................................. 22
   2.4 Annotating data at different levels .................. 28

3 How to collect annotations 29
   3.1 Interfaces for crowdsourcing and task managers ..... 29
   3.2 Labeling task design ................................. 31
   3.3 Evaluating and ensuring quality ........................ 35

4 Which data to annotate 39
   4.1 Active learning ........................................ 39
   4.2 Interactive annotation .................................. 45

5 Conclusions 50

References 52
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.
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 trade-offs 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. 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.
Introduction

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 3.2.3). 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


Full text available at: http://dx.doi.org/10.1561/0600000071


References


References


References


References


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.


Full text available at: http://dx.doi.org/10.1561/0600000071


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, 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.


References


References


