Pattern Recognition Letters 72 (2016) 15–24

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Visipedia circa 2015☆

Serge Belongie[a], Pietro Perona[b,*]

[a] Department of Computer Science, Cornell University and Cornell Tech, United States of America
[b] Department of Electrical Engineering, Department of Computation and Mathematical Sciences, and Computation and Neural Systems Option, California Institute of Technology, United States of America

ARTICLE INFO

Article history:
Available online 10 December 2015

Keywords:
Visipedia
Visual recognition
Human-machine interaction
Machine learning
Active learning
Wikipedia
Visual psychology
Crowdsourcing
Computer Vision

ABSTRACT

Visipedia is a network of people and machines designed to harvest and organize visual information and make it accessible to anyone who has a visual query. We discuss technical challenges arising from Visipedia and discuss their implications for pattern recognition, computer vision, machine learning and visual psychology. Amongst these are discovering visual information that is implicit in experts' brains and in crowds of people and estimating its accuracy. To motivate our thinking we explore a case study, an automated field guide to the birds of North America. We conclude by discussing research directions that are necessary to make progress on Visipedia. An important realisation is that the study of 'computer vision' and 'machine learning' has to be broadened to include the process of information discovery and the dynamic interaction of people and machines in this context. Human-machine systems with no oracle are now within the scope of pattern recognition, machine learning and computer vision.

1. Introduction

In September 2014 Randall Munroe, the author of the popular web comic *xkcd*, published the panel shown in Fig. 1 highlighting "the difference between the easy and the virtually impossible" in Computer Science. The challenge proved irresistible for the computer vision and machine learning team at Flickr who, just one month later, released the "Park or Bird" web app demonstrating that the 'virtually impossible' was, in fact, possible (Fig. 2). Was Munroe proved wrong? Not quite: as he suspected, solving the park question was easy (use the geotag), while solving the bird part was very challenging. The Flickr team's success was enabled by a very recent computer vision and machine learning technical breakthrough [1,2].

As we play with Flickr's clever app, our curiosity is piqued: OK, great, it is a bird. But what *kind* of bird is it? "Bird," like "truck" or "flower," is an *entry level category* [3], at the level of abstraction most people commonly use to think and talk about an object. The species of the bird is called 'subordinate category' by psychologists and 'fine-grained category' by computer vision scientists. We all see the 'bird', but most of us can not recognize its species (we wish we could).

This sentiment, together with mild frustration with current technology, is echoed by Yahoo! president Marissa Mayer (interview in *IEEE Spectrum* in March of 2012):

> You see a bird and want to know what it is. You can take a picture of it, and Google's Goggles will tell you that it's a bird (but you already knew that, didn't you?).

Yes, indeed. Would it not be nice to have an app on our smart device, so that we can point our camera at any object and instantly learn about it in detail? Have we all not wished at some point that we could classify insects, architectural styles, animal bones and many other things? Mayer raises two important points. First, some of our queries are inherently visual. Wikipedia is wonderful, but what do we type into the search box when we wish to identify the bird that is pecking at our birdfeeder? Her second point is that, while there has been much progress in computer vision, the visual queries that are most interesting to humans are not yet addressed by machines.

This brings us to the goal of the Visipedia project, which is *to build a system for discovering and organizing visual information and making it easily accessible to anyone*. We argue that the right solution is similar to Wikipedia, albeit with a bigger role for automation – a decentralized, continually improving collaborative network of people and machines. Obviously, dealing with images is not as easy as dealing with text and thus Visipedia presents a number of unique and interesting challenges in pattern recognition, learning, psychology and beyond. In the remainder of this article we explore some of these challenges and describe our experience tackling a few

☆ This paper has been recommended for acceptance by Rama Chellappa.

* Corresponding author. Tel.: +1 (626) 395-4867.
E-mail addresses: sjb344@cornell.edu (S. Belongie), perona@caltech.edu (P. Perona).

http://dx.doi.org/10.1016/j.patrec.2015.11.023
0167-8655/

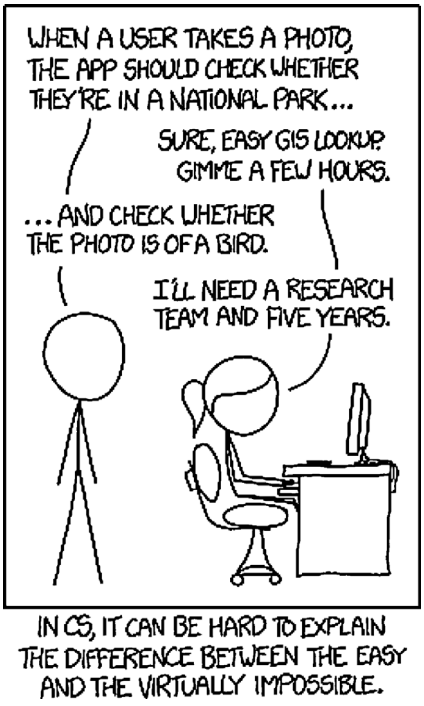

Fig. 1. Comic by Randall Munroe (http://xkcd.com/1425, September 2014).

of them. We will use bird identification as a case study throughout. We conclude with the observation that the scope of pattern recognition, computer vision and machine learning is broader than we thought.

2. A digital field guide for birds

2.1. Merlin, powered by Visipedia

In June 2015 we released an on-line field guide to North American birds called *Merlin*[1]. At the time of writing the system knows 400 out of about 1000 North American bird species and can classify them them from a picture. We engaged in this project in collaboration with the Cornell Laboratory of Ornithology to help us identify the issues that one will encounter in the more general Visipedia setting. Fig. 3 depicts a use case; the screenshots capture the user experience, which starts with photo submission and concludes with a brief description and an audio recording of the identified bird.

The computer vision architecture of *Merlin* consists of four steps [4] reflecting our current understanding of fine-grained classification [4–6]:

1. *Detection* – the bird is localized in the image using a 'bird' detector that is trained to work on all species.
2. *Part registration* – key landmarks (bill, belly, tail, feet, etc.) of a general-purpose 'bird model' are identified in the picture. The corresponding image patches are appropriately rectified.
3. *Feature extraction* – features are computed from the image using a reference frame that is defined by the landmarks.
4. *Fine-grained classification* – the features are used in a multi-class classifier to produce a shortlist of possible matches.

This architecture is quite sophisticated and contains a number of insights on how to represent features, how to compute pose for 3D objects, how to detect categories, etc.

While interesting, the computer vision bits are not the focus of Visipedia. From the point of view of Visipedia this is technology

[1] http://merlin.allaboutbirds.org/photo-id.

Fig. 2. Flickr's response to Munroe (Fig. 1). October 2014 (http://parkorbird.flickr.com).

which, we assume, will continue to be perfected over time by talented computer vision and machine learning researchers. The focus of the Visipedia project is understanding how humans and machines may best cooperate in discovering, organizing and searching visual information.

2.2. Humans and machines are complementary

A first realization we reached through developing and deploying *Merlin* is that the user and the machine have complementary abilities and mutually benefit from collaboration [7,8]. On the one side, the user is ignorant as to bird species and taxonomy (e.g., that the 'pileated woodpecker' is a species of woodpeckers, that it has a red crest, and is related to a number of South American woodpeckers). This is precisely why a user would engage with Visipedia. However, the user can see very well and will detect easily a 'bird' even when half hidden in a bush. He will also detect easily the main parts of the bird: eyes, tail, wings. Conversely, the machine has perfect knowledge of taxonomy and attributes, but cannot see as well as the user. Often (currently about 75% of the time [4]) the machine can detect and classify the bird in the picture. However, if it is confused the machine can ask the user for a hint, e.g., to click on the bill (see Fig. 3), which for the user is an easy and quick task. This simple information is valuable for the machine. It helps part registration (step (2) above), allowing it to read correctly the bird's attributes and complete the task successfully. Thus, *a machine collaborating with a human can solve a visual query that neither human nor machine could solve by themselves*.

Fig. 4 shows screen captures of a collaborative GUI for bird species classification [7,8]. Here a fast and accurate field guide was obtained by combining computer vision algorithms and human observers. Computer algorithms detect parts and predict attributes and human observers answer simple questions (e.g., 'what is the primary color of the bird') or perform simple actions (e.g., 'click on the head'). Building this system involved developing probabilistic models of the strengths and weaknesses of humans and computers for different types of tasks such as predicting part locations, attributes, and classes. The system selects automatically which questions to ask human users – the most informative questions are chosen by maximizing an information gain criterion – this reduces on average the amount of interaction that is needed to achieve a satisfactory answer.

A second realization is that it is not possible to build the bird field guide without help from humans. Computer vision researchers dream of building machines that discover structure in images automatically and that can learn visual categories without supervision [9,10]. However, it is not realistic that all the information necessary to build a field guide will be gathered without the help of human experts. Humans have bodies, explore the world, take things apart. There are things that only humans know. Key to Visipedia's success is enabling humans to share their visual knowledge. It is useful